



Volume 107 February 2015
Number 1

Published quarterly
by the
American Psychological
Association

ISSN 0022-0663

Journal of Educational Psychology

Steve Graham, *Editor*

Jill Fitzgerald, *Associate Editor*

Panayiota Kendeou, *Associate Editor*

Young-Suk Kim, *Associate Editor*

Pui-Wa Lei, *Associate Editor*

Daniel H. Robinson, *Associate Editor*

Cary J. Roseth, *Associate Editor*

Tanya Santangelo, *Associate Editor*

Gregory Schraw, *Associate Editor*

Birgit Spinath, *Associate Editor*

CURRENT YR/VOL

Marygrove College Library
8425 West McNichols Road
Detroit, MI 48221

Editor

Steve Graham, EdD, *Arizona State University*

Associate Editors

Jill Fitzgerald, PhD, *University of North Carolina*
Panayiota Kendeou, PhD, *University of Minnesota*
Young-Suk Kim, EdD, *Florida State University*
Pui-Wa Lei, PhD, *Pennsylvania State University*
Daniel H. Robinson, PhD, *Colorado State University*
Cary J. Roseth, PhD, *Michigan State University*
Tanya Santangelo, PhD, *Arcadia University*
Gregory Schraw, PhD, *University of Nevada, Las Vegas*
Birgit Spinath, PhD, *Heidelberg University*

Consulting Editors

Mary D. Ainley, *University of Melbourne*
Patricia Alexander, *University of Maryland*
Eric Anderman, *The Ohio State University*
Particia Ashton, *University of Florida*
Roderick W. Barron, *University of Guelph*
Matt Bernacki, *University of Nevada, Las Vegas*
David A. Bergin, *University of Missouri*
Daniel Bolt, *University of Wisconsin, Madison*
Mimi Bong, *Korea University*
Lee Branum-Martin, *Georgia State University*
Adriana G. Bus, *Universiteit Leiden*
Kirsten R. Butcher, *University of Utah*
Fabrizio Butera, *University of Lausanne*
Robert Calfee, *Stanford University*
Martha Carr, *University of Georgia*
Becky Xi Chen, *University of Toronto*
Clark Chinn, *Rutgers University*
Tim Cleary, *Rutgers University*
Donald Compton, PhD, *Vanderbilt University*
Pierre Cormier, *Université de Moncton*
Kai Cortina, *University of Michigan*
Michael D. Coyne, Ph.D., *University of Connecticut*
Carol McDonald Connor, *Arizona State University*
Jennifer Cromley, *Temple University*
Anne E. Cunningham, *University of California, Berkeley*
Heather A. Davis, *North Carolina State University*
David K. Dickinson, *Vanderbilt University*
Andrew Elliot, *University of Rochester*
Steve Elliott, *Arizona State University*
Weihsia Fan, *University of Houston*
Ralph Ferretti, *University of Delaware*
Sara J. Finney, *James Madison University*
Brett Foley, *Alpine Testing Solutions*
Barbara Foorman, *Florida State University*
Donna Y. Ford, *Vanderbilt University*
Lynn S. Fuchs, *Vanderbilt University*
David W. Galbraith, *University of Southampton*
Elizabeth Gee, *Arizona State University*
Jim Gee, *Arizona State University*
Michele Gregoire Gill, *University of Central Florida*
Arthur M. Glenberg, *Arizona State University*
Susan Goldman, *University of Illinois*
Art Graesser, *University of Memphis*
Deleon Gray, *North Carolina State University*
Barbara A. Greene, *University of Oklahoma*
Jeffrey A. Greene, *University of North Carolina, Chapel Hill*
Antonio Gutierrez, *Georgia Southern University*
John T. Guthrie, *University of Maryland*
Douglas Hacker, *University of Utah*
Karen Harris, *Arizona State University*
John Hattie, *University of Melbourne*
Karen Rambo-Hernandez, *West Virginia State University*
Flaviu Hodis, *Victoria University of Wellington, New Zealand*
Chris Hulleman, *University of Virginia*
Mina C. Johnson, *Arizona State University*
Nancy Jordan, *University of Delaware*
Malt Joshi, *Texas A&M*
Avi Kaplan, *Temple University*
Carol Anne Kardash, *University of Nevada, Las Vegas*
Noona Kiuru, *University of Jyväskylä, Finland*
Kristin Krajewski, *Justus Liebig Universitaet*
Andy Katayama, *United States Air Force Academy*
Michael J. Kieffer, *New York University*
James S. Kim, *Harvard University*
Paul A. Kirschner, *Open University of the Netherlands*
Robert Klassen, *University of York*
Uta Klusmann, *Leibniz Institute for Science and Mathematics Education*
Beth Kurtz-Costes, *University of North Carolina, Chapel Hill*
Terry Kurz, *Arizona State University*
Hongli Li, *Georgia State University*
Xiaodong Lin, *Columbia University*
Elizabeth A. Linnenbrink-Garcia, *Michigan State University*
Min Liu, *University of Hawaii at Mānoa*
Robert Lorch, *University of Kentucky*
Charles MacArthur, *University of Delaware*
Joseph P. Magliano, *Northern Illinois University*
Scott Marley, *Arizona State University*
Andrew Martin, *University of Sydney, Australia*

Linda Mason, *University of North Carolina, Chapel Hill*

Lucia Mason, *Universita degli Studi di Padova*
Margo A. Mastropieri, PhD, *George Mason University*
Richard E. Mayer, *University of California, Santa Barbara*
Matt McCruden, *Victoria University of Wellington*
Mark McDaniel, *Washington University in St. Louis*
Nicole McNeil, *University of Notre Dame*
David Most, *Colorado State University*
P. Karen Murphy, *The Pennsylvania State University*

Benjamin Nagengast, *Eberhard Karls University of Tübingen*
John Nietfeld, *North Carolina State University*
Nikos Ntoumanis, *University of Birmingham*
E. Michael Nussbaum, *University of Nevada, Las Vegas*
Rollanda E. O'Connor, *University of California, Riverside*
Tenaha O'Reilly, *Educational Testing Service*
Fred Paas, *Erasmus University*
Erika Patall, *The University of Texas at Austin*
Helen Patrick, *Purdue University*
Reinhard Pekrun, *University of Munich*
Yaakov Petscher, *Florida State University*
Gary Phye, *Iowa State University*
Pablo Pirnay-Dumma, *Martin-Luther-Universität Halle-Wittenberg, Halle, Germany*
Jan L. Plass, *New York University*
Patrick Proctor, *Boston College*

David Rapp, *Northwestern University*
Katherine Rawson, *Kent State University*
Alexander Renkl, *University of Freiburg*
Lindsey Richland, *University of Chicago*
Gert Rijlaarsdam, *Universiteit van Amsterdam*
Gregory Roberts, *The University of Texas at Austin*
Alysia D. Roehrig, *Florida State University*
Christopher A. Sanchez, *Oregon State University*
Dale Schunk, *University of North Carolina, Greensboro*
Timothy Shanahan, *University of Illinois, Chicago*
Gale M. Sinatra, *University of Southern California*
Susan Sonnenschein, *University of Maryland Baltimore*

The main purpose of the *Journal of Educational Psychology*® is to publish original, primary psychological research pertaining to education across all ages and educational levels. A secondary purpose of the *Journal* is the occasional publication of exceptionally important theoretical and review articles that are pertinent to educational psychology.

Single Issues, Back Issues, and Back Volumes: For information regarding single issues, back issues, or back volumes, write to Order Department, American Psychological Association, 750 First Street, NE, Washington, DC 20002-4242; call 202-336-5600 or 800-374-2721; or visit www.apa.org/pubs/journals/subscriptions.aspx

Manuscripts: Submit manuscripts electronically through the Manuscript Submissions Portal found at www.apa.org/pubs/journals/edu according to the Instructions to Authors found elsewhere in this issue (see table of contents). Correspondence regarding manuscripts should be sent to the Incoming Editor, Steve Graham, at steve.graham@asu.edu. The opinions and statements published are the responsibility of the authors, and such opinions and statements do not necessarily represent the policies of APA or the views of the Editor.

Copyright and Permission: Those who wish to reuse APA-copyrighted material in a non-APA publication must secure from APA written permission to reproduce a journal article in full or journal text of more than 800 cumulative words or more than 3 tables and/or figures. APA normally grants permission contingent on permission of the author, inclusion of the APA copyright notice on the first page of reproduced material, and payment of a fee of \$25 per page. Libraries are permitted to photocopy beyond the limits of the U.S. copyright law: (1) post-1977 articles, provided the per-copy fee in the code for this journal (0022-0663/15/\$12.00) is paid through the Copyright Clearance Center, 222 Rosewood Drive, Danvers, MA 01923; (2) pre-1978 articles, provided that the per-copy fee stated in the Publishers' Fee List is paid through the Copyright Clearance Center. For more information along with a permission form, go to www.apa.org/about/contact/copyright/index.aspx

Electronic Access: APA members who subscribe to this journal have automatic access to all issues of the journal in the *PsycARTICLES*® full-text database. See <http://my.apa.org/access.html>.

Reprints: Authors may order reprints of their articles from the printer when they receive proofs.

APA Journal Staff: Susan J. A. Harris, *Senior Director, Journals Program*; John Breithaupt, *Director, Journal Production Services*; Stephanie Pollock, *Managing Director*; Katie Einhorn, *Account Manager*; Jodi Ashcraft, *Director, Advertising Sales and Exhibits*.

Journal of Educational Psychology® (ISSN 0022-0663) is published quarterly (February, May, August, November) in one volume per year by the American Psychological Association, 750 First Street, NE, Washington, DC 20002-4242. Subscriptions are available on a calendar year basis only (January through December). The 2015 rates follow: *Nonmember Individual*: \$229 Domestic, \$258 Foreign, \$271 Air Mail. *Institutional*: \$821 Domestic, \$870 Foreign, \$885 Air Mail. *APA Member*: \$99. *APA Student Affiliate*: \$69. Write to Subscriptions Department, American Psychological Association, 750 First Street, NE, Washington, DC 20002-4242. Printed in the U.S.A. Periodicals postage paid at Washington, DC, and at additional mailing offices. **POSTMASTER:** Send address changes to *Journal of Educational Psychology*, American Psychological Association, 750 First Street, NE, Washington, DC 20002-4242.

Educational Psychology[®]

www.apa.org/pubs/journals/edu

February 2015

Volume 107
Number 1

Articles

© 2015
American
Psychological
Association

1 Inaugural Editorial for the *Journal of Educational Psychology*
Steve Graham

3 Editorial
Art Graesser

4 Important Text Characteristics for Early-Grades Text Complexity
Jill Fitzgerald, Jeff Elmore, Heather Koons, Elfrieda H. Hiebert, Kimberly Bowen, Eleanor E. Sanford-Moore, and A. Jackson Stenner

30 Successful Learning With Multiple Graphical Representations and Self-Explanation Prompts
Martina A. Rau, Vincent Aleven, and Nikol Rummel

47 An Imagination Effect in Learning From Scientific Text
Claudia Leopold and Richard E. Mayer

64 Matching Learning Style to Instructional Method: Effects on Comprehension
Beth A. Rogowsky, Barbara M. Calhoun, and Paula Tallal

79 Toward an Understanding of Dimensions, Predictors, and the Gender Gap in Written Composition
Young-Suk Kim, Stephanie Al Otaiba, Jeanne Wanzek, and Brandy Gatlin

96 Cross-Language Transfer of Word Reading Accuracy and Word Reading Fluency in Spanish–English and Chinese–English Bilinguals: Script-Universal and Script-Specific Processes
Adrian Pasquarella, Xi Chen, Alexandra Gottardo, and Esther Geva

111 Bilingual Phonological Awareness: Reexamining the Evidence for Relations Within and Across Languages
Lee Branum-Martin, Sha Tao, and Sarah Garnaat

126 Literacy Skill Development of Children With Familial Risk for Dyslexia Through Grades 2, 3, and 8
Kenneth Eklund, Minna Torppa, Mikko Aro, Paavo H. T. Leppänen, and Heikki Lyytinen

141 Parallel and Serial Reading Processes in Children’s Word and Nonword Reading
Madelon van den Boer and Peter F. de Jong

152 Classmate Characteristics and Student Achievement in 33 Countries: Classmates’ Past Achievement, Family Socioeconomic Status, Educational Resources, and Attitudes Toward Reading
Ming Ming Chiu and Bonnie Wing-Yin Chow

170 To What Extent Do Teacher–Student Interaction Quality and Student Gender Contribute to Fifth Graders’ Engagement in Mathematics Learning?
Sara E. Rimm-Kaufman, Alison E. Baroody, Ross A. A. Larsen, Timothy W. Curby, and Tashia Abry

(Contents continue)

186	"Michael Can't Read!" Teachers' Gender Stereotypes and Boys' Reading Self-Concept <i>Jan Retelsdorf, Katja Schwartz, and Frank Asbrock</i>
195	Gender Differences in the Effects of a Utility-Value Intervention to Help Parents Motivate Adolescents in Mathematics and Science <i>Christopher S. Rozek, Janet S. Hyde, Ryan C. Svoboda, Chris S. Hulleman, and Judith M. Harackiewicz</i>
207	Prekindergarten Children's Executive Functioning Skills and Achievement Gains: The Utility of Direct Assessments and Teacher Ratings <i>Mary Wagner Fuhs, Dale Clark Farran, and Kimberly Turner Nesbitt</i>
222	The Effect of Training and Consultation Condition on Teachers' Self-Reported Likelihood of Adoption of a Daily Report Card <i>Alex S. Holdaway and Julie Sarno Owens</i>
236	Earlier School Start Times as a Risk Factor for Poor School Performance: An Examination of Public Elementary Schools in the Commonwealth of Kentucky <i>Peggy S. Keller, Olivia A. Smith, Lauren R. Gilbert, Shuang Bi, Eric A. Haak, and Joseph A. Buckhalt</i>
246	Developmental Dynamics Between Children's Externalizing Problems, Task-Avoidant Behavior, and Academic Performance in Early School Years: A 4-Year Follow-Up <i>Riitta-Leena Metsäpelto, Eija Pakarinen, Noona Kiuru, Anna-Maija Poikkeus, Marja-Kristiina Lerkkanen, and Jari-Erik Nurmi</i>
258	The Big-Fish-Little-Pond Effect: Generalizability of Social Comparison Processes Over Two Age Cohorts From Western, Asian, and Middle Eastern Islamic Countries <i>Herbert W. Marsh, Adel Salah Abduljabbar, Alexandre J. S. Morin, Philip Parker, Faisal Abdelfattah, Benjamin Nagengast, and Maher M. Abu-Hilal</i>
272	Social Consequences of Academic Teaming in Middle School: The Influence of Shared Course Taking on Peer Victimization <i>Leslie Echols</i>
284	Long-Term Implications of Early Education and Care Programs for Australian Children <i>Rebekah Levine Coley, Caitlin McPherran Lombardi, and Jacqueline Sims</i>
300	"He Who Can, Does; He Who Cannot, Teaches?": Stereotype Threat and Preservice Teachers <i>Toni A. Ihme and Jens Möller</i>
309	Value Development Underlies the Benefits of Parents' Involvement in Children's Learning: A Longitudinal Investigation in the United States and China <i>Cecilia Sin-Sze Cheung and Eva M. Pomerantz</i>

Other

320	E-Mail Notification of Your Latest Issue Online!
iii	Instructions to Authors
78	Members of Underrepresented Groups: Reviewers for Journal Manuscripts Wanted
46	Subscription Order Form

With the exception of the editorials, all articles in this issue were accepted during the editorial term of Art C. Grasser.

EDITORIAL

Inaugural Editorial for the *Journal of Educational Psychology*

My editorship of the *Journal of Educational Psychology* (*JEP*) begins with this issue. I am excited and pleased to start this journey, as I believe that *JEP* is without peer in the educational research world. It publishes the most important, highest quality research in educational psychology and education more broadly. As I begin this venture, I am acutely aware that the job of an editor is not an easy one. As the poet John Wheelock noted, it is the “dullest, hardest, most exciting, exasperating and rewarding of perhaps any job in the world” (Charlton, 1997, p. 142). Editors play a pivotal role in the research enterprise, vetting, shaping, and improving the presentation of the work submitted to them, but it is an arduous process that involves peer review, agonizing decisions about what to publish, and painstaking attention to detail.

Fortunately, I am not making this journey alone. I am joined by an outstanding group of associate editors that includes (in alphabetical order) Jill Fitzgerald, Pani Kendeou, Pui-Wa Lei, Dan Robinson, Cary Rose, Tanya Santangelo, Gregg Shrav, Birgit Spinath, and Young Suk-Kim. We are joined by a highly talented, diverse, and international board of consulting editors and principal reviewers. Principal reviewers are a new addition to *JEP*. They make a welcomed commitment to serve the journal by reviewing between four and six manuscripts a year. Their efforts are recognized in the last issue of each volume year.

One of my favorite comments about editors comes from Samuel Clemens, who quipped, “How often we recall, with regret, that Napoleon once shot at a magazine editor and missed him and killed a publisher. But we remember with charity that his intentions were good” (Ayres, 1997, p. 66). At one time or another, I suspect all of us have harbored a negative thought or two about editors. Although there are many possible reasons for this, one of the most exasperating involves an inordinately long review process. We on the *JEP* editorial team are committed to making sure this does not happen here. Our goal is for authors to receive a decision on their manuscript, based on a sound evaluation by their peers, in a timely manner—in 90 days or less, with an emphasis on *less*. If a paper is clearly not appropriate for *JEP*, we will let authors know why immediately.

As the new editor of *JEP*, I am acutely aware of my responsibilities to the journal, educational psychology, and the field of education in general. My predecessors comprise a formidable array of talent and editorial wizardry, including Raymond Kuhlen, Wayne Holtzman, Johanna Williams, Samuel Ball, Robert Calfee, Joel Levin, Michael Pressley, Karen Harris, and Art Graesser. To quote a former editor, they made *JEP* the leading “outlet in the world for psychologically oriented research in education” (Pressley, 1997, p. 3). They accomplished this by publishing high-quality investigations, articles that moved the field forward conceptually and empirically, and the strongest interdisciplinary research in education. They encouraged others to submit their best work to the journal and made hard decisions about what to publish. We will uphold these traditions.

This does not mean that maintaining the status quo is our objective. We plan to make *JEP* even better. How do we plan to achieve this goal? At the most basic level, we want to make sure the work submitted and published in the journal is as good as it can be. This means that before a study is reviewed for *JEP*, it must meet certain criteria. Before submitting an article to *JEP*, we encourage authors to examine the Journal Article Reporting Standards specified in Volume 63 of the 2008 *American Psychologist* (APA Publications and Communications Board Working Group on Journal Article Reporting Standards, 2008) or the appendix in the *Publication Manual of the American Psychological Association* (American Psychological Association, 2010).

Some criteria must be met before we send a paper out for review. First, the participants and the setting in which the research occurred must be adequately described. Such descriptions are essential to contextualizing and interpreting the findings from a study, replicating an investigation, determining generalizability of findings, and conducting meta-analyses.

Second, there must be adequate evidence that measures are reliable and valid. Because reliability is a characteristic of the sample and the measure (Crocker & Algina, 1986; Harris, 2003), referencing previous evidence to support reliability is not enough in many cases. In these instances, authors must provide evidence that the measures used in their study were reliable with their sample.

A sometimes vexing set of measures in terms of reliability and validity are grades and grade point averages. In some instances, grades are based on one or more reliable exams administered to all participants, but too often they are based on undefined and likely different procedures that vary by the class or classes students completed. Although grades and grade point averages are a legitimate area of study in their own right, we are

reluctant to review a paper when the reliability and validity of grades is questionable, especially if they serve as the only outcome measures of achievement or academic progress in a study. Before such a study is reviewed, it is incumbent on the authors to provide convincing evidence that these measures are, in fact, adequate.

Third, we would like for *JEP* to publish even more intervention research than it does now, but before we review an intervention study, authors must provide evidence that the treatment was implemented as intended. Simply put, trust cannot be placed in findings from a study in which treatment validity was not established. Such evidence is essential to any claim that a specific intervention was responsible for observed changes. Of course, it is equally important to describe what happened in control and comparison conditions.

Fourth, authors may be asked to upgrade their statistical analyses before we review a paper. We hope this will not occur often, but we are especially sensitive to this issue for large as well as longitudinal databases. The process of raising questions and concerns about data analyses commonly occurs during the peer review process, but if there is an evident issue, we will ask for it to be resolved before the paper is reviewed.

As noted earlier, one of our intents in implementing these criteria is to shape and enhance the quality of work submitted and published in *JEP*. Two other important purposes are served as well. First, authors increase the probability of receiving a positive review when participants and setting are adequately described, measures are reliable and valid, treatment fidelity is established, and appropriate statistical procedures are applied. Second, reviewers do not spend valuable time reviewing a manuscript missing fundamental information.

We further require authors submitting manuscripts to *JEP* to report appropriate effect sizes as well as means, standard deviations, and confidence intervals for their variables. More than 10 years ago, Harris (2003) made a similar call, but we still receive a sizable number of submissions missing such basic data.

Another way we plan to make *JEP* even better is to communicate to the field our interest in publishing high-quality research involving multiple methodologies, including quantitative, qualitative, single-subject, and mixed-methods designs. This interest extends to other forms of scholarship, especially meta-analyses, but includes conceptual, methodological, and integrative reviews of the literature too. The world of educational psychology is very diverse in its interest and approaches to scholarship. We hope that during our watch, *JEP* can become even better at capturing this complexity.

For an editor, it is a bit risky to specify the types of papers you do not plan to publish, as exceptions may be made along the way. With that said, we do not plan to publish survey studies that are based on convenience or unrepresentative samples of respondents. Nor do we plan to publish studies that primarily focus on creating and validating a test or specific measures. Although such studies are important and needed, a variety of journals serve this purpose. Finally, replication studies are important to science and the field. Replications of new findings contained in a single article are encouraged. The journal might also look favorably on multiple replications of a prior study in a submitted paper. For the most part, though, a new study needs to do more than simply replicate a previously published investigation. It needs to make an important extension to understanding of the phenomena under investigation. Thus, systematic replications that both reproduce and extend the original research are encouraged.

As I bring this editorial to a close, I want to indicate how pleased I am that *JEP* has become so international. This has become increasingly evident over the last 20 years in terms of editorial board members, submissions, and publications. This is a trend I and my team plan to nurture. We further invite academics who are interested in reviewing for the journal to contact me (steve.graham@asu.edu): Send your vita and tell me about your expertise. In the best interest of student training, we encourage reviewers to invite doctoral students to complete reviews with them. We acknowledge these students' contribution in the final issue of each volume. Such apprenticeships are essential to growing a healthy field.

In closing, I have one last thought to share with you. If you send us a paper and we publish it, please accept our thanks and send us more papers! If we do not publish it, the same sentiment applies.

Steve Graham, Editor.

References

American Psychological Association. (2010). *Publication manual of the American Psychological Association* (6th ed.). Washington, DC: Author.

APA Publications and Communications Board Working Group on Journal Article Reporting Standards. (2008). Reporting standards for research in psychology: Why do we need them? What might they be? *American Psychologist*, 63, 839–851.

Ayres, A. (Ed.). (2005). *The wit and wisdom of Mark Twain*. New York, NY: HarperCollins.

Charlton, J. (Ed.). (1997). *The writer's quotation book: A literary companion* (4th ed.). Boston, MA: Faber & Faber.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Orlando, FL: Harcourt Brace Jovanovich.

Harris, K. R. (2003). Editorial: Is the work as good as it could be? *Journal of Educational Psychology*, 95, 451–452. doi:10.1037/0022-0663.95.3.451

Pressley, M. (1997). Editorial. *Journal of Educational Psychology*, 89, 3–4. doi:10.1037/h0092689

Editorial

I accepted the editorship of *Journal of Educational Psychology* in 2008 with the hopes of achieving several goals. Concrete missions motivate an editor to devote the necessary time and energy to editing over a nontrivial span of one's career. Now that I have completed my 6-year term as editor, I can reflect on the extent to which my goals were achieved.

The first, rather obvious, goal was to have the journal grow with an abundance of high-quality research. The number of submissions indeed grew by approximately 50% to 550 new submissions per year and the pages increased by 33% to 1,200 printed pages per volume. During the same period, the rejection rate increased to 83%. Thanks to the 11 associate editors, over 100 colleagues on the editorial board, and thousands of ad hoc reviewers, the quality of the reviews maintained the historically high standards of this journal. Of course, these numbers are reassuring, but what about the more substantive goals?

The second goal was to encourage studies with objective measures of learning, achievement, social interaction, motivation, emotion, and other psychological constructs relevant to education. Objective measures are grounded in behavior, performance, cognitive tasks, objective tests, and neuroscience. Psychological rating scales and other forms of self-report are important sources of data when combined with objective measures, but exclusive reliance on self-report data is a flimsy foundation for science in the 21st century. The associate editors shared this perspective and applied rigorous measurement standards in the review process. Our perspective did disappoint some authors who had built a cottage industry of administering psychological tests with self-report measures to samples of participants who were available (convenience samples rather than representative samples) and reporting correlations among these self-report measures (typically without replication). However, serious complaints about our position were surprisingly rare and hopefully reflected general improvements in measurement standards in the field of educational psychology.

A third goal was to increase coverage of research with computer technologies. Educational technologies have had a revolutionary effect on education during the last decade; therefore, it is important for the journal to capture those trends. Computers can reliably collect objective and self-report measures at a fine-grain level, systematically implement pedagogical interventions, and quickly analyze data. The journal did have an increase in publications with educational technologies over the course of my editorship. This was partly reflected in a special issue on advanced learning technologies (the only special issue under my editorship). The hope is that this journal continues to encourage submissions with computer technologies, such as multimedia, intelligent tutoring systems, conversational agents, social media, educational games, distributed learning environments, and conventional computer-based training.

A fourth goal was to encourage studies that coordinate educational data mining methodologies with theory-based evidence-centered design and measures that satisfy psychometric standards. For example, there was a special section of an issue that focused on analyzing computer logs in large-scale assessments. Hundreds of observations per hour can be tracked by computers, including response times. Such rich data can be mined to discover patterns of data that might not be anticipated by researchers *a priori*, but they can be linked to psychological constructs and thereby advance educational theory. Progress on this fourth goal did not progress to my satisfaction, but it will hopefully evolve in future years.

The most difficult challenge as editor was in finding ways to handle a large number of manuscripts with sophisticated quantitative techniques that stretched beyond the conventional analysis of variance, multiple regression, and nonparametric statistical analyses. There were not enough colleagues with expertise in advanced statistics to handle the load. Indeed, universities are not graduating a sufficient number of doctoral degrees in quantitative areas of the social sciences to meet the demands of several areas of psychology. The challenge was compounded by the fact that many of the manuscripts with advanced statistical techniques were quite lengthy; therefore, colleagues were prone to decline reviewing them (even after they initially agreed to review the manuscripts). Consequently, some manuscripts required more than 3 months for review and there was high turnover in associate editors with sophisticated quantitative expertise. We need to find ways to fill the serious expertise gap in advanced research designs and statistics.

In closing, I would like to thank Jean Edgar, my chief editorial assistant. She assisted me with compassion and enthusiasm for over 6 years.

Art Graesser, Editor

Important Text Characteristics for Early-Grades Text Complexity

Jill Fitzgerald

MetaMetrics, Durham, North Carolina, and The University of North Carolina at Chapel Hill

Heather Koons

MetaMetrics, Durham, North Carolina, and The University of North Carolina at Chapel Hill

Kimberly Bowen and Eleanor E. Sanford-Moore

MetaMetrics, Durham, North Carolina

Jeff Elmore

MetaMetrics, Durham, North Carolina

Elfrieda H. Hiebert

TextProject, Santa Cruz, California and The University of California at Santa Cruz

A. Jackson Stenner

MetaMetrics, Durham, North Carolina, and The University of North Carolina at Chapel Hill

The Common Core set a standard for all children to read increasingly complex texts throughout schooling. The purpose of the present study was to explore text characteristics specifically in relation to early-grades text complexity. Three hundred fifty primary-grades texts were selected and digitized. Twenty-two text characteristics were identified at 4 linguistic levels, and multiple computerized operationalizations were created for each of the 22 text characteristics. A researcher-devised text-complexity outcome measure was based on teacher judgment of text complexity in the 350 texts as well as on student judgment of text complexity as gauged by their responses in a maze task for a subset of the 350 texts. Analyses were conducted using a logical analytical progression typically used in machine-learning research. Random forest regression was the primary statistical modeling technique. Nine text characteristics were most important for early-grades text complexity including word structure (decoding demand and number of syllables in words), word meaning (age of acquisition, abstractness, and word rareness), and sentence and discourse-level characteristics (intersentential complexity, phrase diversity, text density/information load, and noncompressibility). Notably, interplay among text characteristics was important to explanation of text complexity, particularly for subsets of texts.

Keywords: text complexity, early-grades reading, random forest regression, machine-learning

Supplemental materials: <http://dx.doi.org/10.1037/a0037289.supp>

The United States Common Core State Standards (CCSS) for English Language Arts (National Governors Associate Center for Best Practices [NGA] & Council of Chief State School Officers

[CCSSO], 2010) bring unprecedented attention to the nature of texts that students read. The goal of the Standards is for high school graduates to be well prepared for college and workplace careers. The ability to read college-and-workplace texts plays a prominent role in the Standards for that preparation. Citing prior evidence of a current-day gap between the text-complexity levels at high-school graduation and college and workplace (e.g., ACT, 2006; Williamson, 2008), the CCSS authors set a challenging standard for *all* students to be able to “comprehend texts of steadily increasing complexity as they progress through school . . .” (NGA & CCSSO, 2010, Appendix A, p. 2). The foundation for students’ ability to read increasingly complex texts begins in early reading exposure, and considerable controversy and debate has focused attention on the potential impact of the text-complexity Standard for young readers (e.g., Hiebert, 2012; Messmer, Cunningham, & Hiebert, 2012). As educators attempt to support youngsters to read increasingly complex texts, early-grades teachers need a sound understanding of what makes texts more or less complex for young students who are beginning to learn to read. An empirically based understanding of text complexity for early-grades readers is critical for practical reasons and should also contribute to development of theoretical modeling of

This article was published Online First August 4, 2014.

Jill Fitzgerald, MetaMetrics, Durham, North Carolina, and School of Education, The University of North Carolina at Chapel Hill; Jeff Elmore, MetaMetrics, Durham, North Carolina; Heather Koons, MetaMetrics, Durham, North Carolina, and School of Education, The University of North Carolina at Chapel Hill; Elfrieda H. Hiebert, TextProject, Santa Cruz, California, and Department of Education, The University of California at Santa Cruz; Kimberly Bowen and Eleanor E. Sanford-Moore, MetaMetrics, Durham, North Carolina; A. Jackson Stenner, MetaMetrics, Durham, North Carolina, and School of Education, The University of North Carolina at Chapel Hill.

The authors wish to disclose that Jill Fitzgerald, Jeff Elmore, Heather Koons, Kimberly Bowen, Eleanor E. Sanford-Moore, and A. Jackson Stenner are employees of MetaMetrics, Inc. Elfrieda H. Hiebert is a consultant at MetaMetrics, Inc. MetaMetrics, Inc. funded this project.

Correspondence concerning this article should be addressed to Jill Fitzgerald, 8565 Nicholson Farm Lane, Graham, NC 27253. E-mail: jfitzger@email.unc.edu

text complexity. The purpose of the present study was to explore text characteristics specifically in relation to early-grades text complexity. The research questions addressed in the study were as follows: (a) Which text characteristics are most important for early-grades text complexity? (b) Is there interplay of text characteristics in relation to text complexity, and if there is, can any aspects of the interplay be described? The research questions were addressed using computer-based analysis of texts. The present study makes an additional contribution to the educational research literature in that a statistical approach and methodological sequence unique in the educational research literature were used—random forest regression in conjunction with a machine-learning research paradigm.

What Is Text Complexity?

On a broad stage, in science writ large, “complexity” has overtaken “parsimony” as a focal interest in both physical and social sciences. Scientists increasingly aim to understand complexity as it exists naturally in the world—as opposed to more traditional efforts to reduce natural occurrences to some fundamental simplicity (e.g., Bar-Yam, 1997). The seminal philosophical definition of complexity may be attributed to Rescher (1998, p. 1)—“Complexity is . . . a matter of the number and variety of an item’s constituent elements and of the elaborateness of their interrelational structure, be it organizational or operational.” Complexity theory suggests that although the complexity of some objects, events, or actions may not be fully understood, three essential elements of complex systems can be pinpointed and characterized (Bar-Yam, 1997; Kauffman, 1995). First, in general, complex systems involve a large number of mutually interacting parts, but even a small number of interacting components can behave in complex ways (Albert & Barabási, 2002; Bar-Yam, 1997). When complexity occurs, a reciprocal relationship exists between parts and wholes. Ensembles are influenced by the distinct elements, but the distinct elements are also influenced by the whole of the ensemble (Merlini Barbaresi, 2003). Second, however, there is usually a limit to the number of parts the researcher has primary interest in, and paradoxically, for practical and research purposes, often summative description of a complicated system may require description as a particular few-part system where the few-part system retains the character of the whole (Bar-Yam, 1997). Third, most complex systems are purposive, and there is often a sense in which the systems are engineered (Bar-Yam, 1997).

Following suit, for the present study, a dynamic systems definition of text complexity was embraced. First, “text” is defined as “. . . an organized unit, whose various components or levels are recognized to give autonomous contributions to the global effect . . .” (Merlini Barbaresi, 2002, p. 120). Second, text complexity is “. . . a dynamic configuration resulting from the contributions of complex phenomena, as they occur at the various text levels” and across text levels (Merlini Barbaresi, 2003, p. 23). The CCSS text-complexity definition further undergirded the present work—text complexity is “the inherent difficulty of reading and comprehending text combined with consideration of the reader and task variables” (NGA & CCSSO, 2010, Appendix A, Glossary of Key Terms, p. 43). The Common Core definition is embedded in a systems outlook in which complexity arises among reader, printed text, and situation during the whole of a reading act. That is, when

engaged in a specific reading encounter, complexity is in some degree relative to an individual and to contextual characteristics (such as age or developmental reading level or degree of teacher support while reading). Concomitantly, complexity of particular texts is relative to populations of readers at different ages or reading ability levels (cf. Kusters, 2008, and Miestamo, 2009, on relative vs. absolute complexity; van der Sluis & van den Broek, 2010). That is, when viewed on a continuum of complexity in relation to many readers’ developmental levels, texts have an emergent nature and can be assigned a “complexity level” to situate them on an entire continuum. The stance is consistent with theories of reading dating back to Rosenblatt’s expositions on reading as transactional (Rosenblatt, 1938, 2005) and Rumelhart’s (1985) explanation of reading as interactive and, more recently, to the widely accepted Rand Reading Study Group model of reading (Snow, 2002). For example, in the Rand Reading Study Group model, text is squarely rooted in an interaction with the reader as reading happens during an activity within a particular social context. The stance is also consistent with Mesmer et al.’s (2012) exposition of early-grades text characteristics in that they also address text complexity as situated within individual and social/instructional contexts.

Commensurate with the three essential elements named above for complex systems, for the present study, we assumed (a) that early-grades texts are complex systems consisting of many mutually interacting characteristics and ensembles of characteristics that interplay to impact text complexity, and the characteristics can be quantitatively measured; (b) to begin to understand the text-characteristic functioning, we would need to consider an organizational scheme for the characteristics and explore whether and how characteristics interact; and (c) the complexity of early-grades texts purposefully exists (i.e., it is in some sense engineered) to support young children to learn to read with as much ease as possible. As well, exploration of interplay among text characteristics would be essential to successful explanation of text complexity.

Which Text Characteristics Might Matter Most for Early-Grades Text Complexity?

An “optimal” text is one in which text characteristics are configured such that readers can construct meaning while engaged with the text with the greatest amount of ease *and* the greatest depth of processing (cf. Merlini Barbaresi, 2003, on optimality theory and Juola, 2003, on the necessity of complex systems to reflect “process,” including cognitive process). Text authors may consciously or unconsciously use optimality when creating texts for particular audiences. Generally, authors must make trade-off choices between favoring readers’ processing ease (efficiency) and readers’ processing depth (effectiveness), and the point of balance between the two is constrained by intended uses of the text, including intended readers of the text (cf. Merlini Barbaresi, 2003, who references the trade-offs, but in recognition of how an author develops a text, rather than in reference to readers/audience). For example, in content-laden disciplinary texts, readers’ processing depth (effectiveness) is often given preference over readers’ processing ease (efficiency). Early-grades texts are generally created to heighten certain factors related to children’s processing ease (such as word decodability), while simultaneously requiring a

relatively low level of processing depth, that is, requiring little effort for meaning creation. Further, some evidence suggests that text characteristics do influence the early word-reading strategies that young children develop (Compton, Appleton, & Hosp, 2004; Juel & Roper-Schneider, 1985). For example, in one study, when tested on novel words, young students who read highly decodable texts outperformed other students who primarily read texts with repetition of high-frequency words (Juel & Roper-Schneider, 1985).

The concept of optimality suggests that different text characteristics might be more important at certain levels of students' reading development than at others, leading directly to consideration of which characteristics of text might be related to the development of students' emergent reading ability. A deep research base suggests that, while meaning creation is at the heart of learning to read, "cracking the code" requires focal effort for beginning readers, and critical cognitive factors inherent in the early learning-to-read phase are development of phonological awareness and word recognition (e.g., Adams, 1990; Fitzgerald & Shanahan, 2000). As a result, hypothetical critical text characteristics that would support early word-reading development are, for example, texts that are composed of: repetition of simple words, which likely facilitates sight word development and orthographic-pattern knowledge (e.g., Metsala, 1999; Vadasy, Sanders, & Peyton, 2005); words with relatively simple orthographic configurations, which facilitates orthographic-pattern knowledge (e.g., Bowers & Wolf, 1993); rhyming words, which may advance phonological awareness (e.g., Adams, 1990); words that are familiar in meaning in oral language, which likely reduce challenges to meaning creation while reading, permitting more attention to word recognition (e.g., Muter, Hulme, Snowling, & Stevenson, 2004); and repeated refrains or repetitive phrases, which likely reinforce phonological awareness and development of sight words along with varied word recognition strategies such as using context to make guesses at unknown words (e.g., Ehri & McCormick, 1998; cf. Bazzanella, 2011, on multiple functions of repetition in oral discourse, including cognitive facilitation). Moreover, inclusion of several types of text-characteristic support might exponentially boost students' ease of learning about code-related facets of reading.

Consequently, to describe early-grades text complexity, it is theoretically necessary to consider several text characteristics at multiple linguistic levels (Graesser & McNamara, 2011; Graesser, McNamara, & Kulikowich, 2011; Kintsch, 1998; Snow, 2002). Studying linguistic levels in text complexity is compatible with research that suggests that hierarchy is one of the central architectures of complexity (Simon, 1962). The research base supporting the importance of multiple levels of texts characteristics for early phases of learning to read is extensive and comprehensive (Mesmer et al., 2012). Only illustrative citations are provided in the following summary (which compares to Mesmer et al., 2012).

Beginning readers learn to attach specific sounds to graphemes and vice versa (e.g., Fitzgerald & Shanahan, 2000), and the research base on the importance of phonological activity is extensive (e.g., Schatschneider, Fletcher, Francis, Carlson, & Foorman, 2004). Other aspects of word-level features have also received wide attention in early-grades texts. In particular, word structure (how a word is configured) and word frequency (the degree to which a word occurs in spoken or written language) have deep research bases. With regard to word structure, letter-sound regu-

larity in words is highlighted in decodable and linguistic texts where significant attention is paid to word rimes and bigrams and trigrams (two and three letter units). Such texts have been shown to have positive impact on oral reading accuracy, but not on comprehension or other global measures of reading (e.g., Compton et al., 2004). With regard to word familiarity, many early grades texts are designed to include repetition of high-frequency words. Children's accuracy and speed of recognition is influenced by word frequency (e.g., Howes & Solomon, 1951).

The importance of knowing key meanings in texts has been well substantiated in relation to its impact on comprehension (e.g., Stanovich, 1986), and some evidence suggests that young students may benefit from texts with easier and more familiar vocabulary (e.g., Hiebert & Fisher, 2007). However, current-day early-grades texts may contain a fairly large amount of challenging word meanings (e.g., Foorman, Francis, Davidson, Harm, & Griffin, 2004). In general, words that occur with higher frequency are processed more quickly and tend to be associated with networks of knowledge (Graesser et al., 2011). In addition to word frequency, other word meaning factors, including imageability, concreteness, and age of word acquisition, have been shown to be significant for students' comprehension and/or word recognition during reading (e.g., Woolams, 2005).

Within-sentence syntax is primarily related to the ease or challenge for creating meaning while reading, as opposed to word recognition (Mesmer et al., 2012). The importance of within-sentence syntax in texts is likely due to the extent to which complexity within a sentence places demands on children's working memory (Graesser et al., 2011).

Discourse-level text characteristics impact aspects of reading in general (Graesser et al., 2011) and are likely to be related to early reading. For example, referential cohesion—occasions when a noun, pronoun, or noun phrase reference another element in the text—has been shown to be related to reading time and comprehension (e.g., McNamara & Kintsch, 1996). More cohesive texts tend to facilitate comprehension, likely because they support mental model building (Kintsch, 1998). It has long been known that even young readers have expectations for story structures that they tend to use to guide comprehension, although young students tend to reveal such expectations to a lesser extent than do older students (e.g., Mandler & Johnson, 1977; Whaley, 1981). As well, better readers make use of informational text structures for comprehension and recall (Britton, Glynn, Meyer, & Penland, 1982). A final potential discourse-level text characteristic is genre, generally considered by linguists and discourse analysts to be a slippery construct (Rudrum, 2005; Steen, 1999). However, questions remain about the relationship between genres and text complexity, especially with regard to identification of various genres according to specific text features (e.g., Mesmer et al., 2012). For instance, findings on the view that narratives are easier texts than other genres are mixed (e.g., Langer et al., 1995, supported the view, while Duke, 2000, did not).

In addition to considering which sorts of text characteristics might be especially important for examining early-grades text complexity, it is essential to embrace potential interplay among various text characteristics. Theoretically, the emergent nature of text complexity is in part due to the challenge level of the constituent elements, but it may also develop through the interplay of the elements (Merlini Barbaresi, 2003). Complex systems tend to have

subsystems that may conflict depending on their “targets,” and to attain a successful result, subsystems need to co-operate toward a compromise solution (Merlini Barbaresi, 2003; cf. Gamson, Lu, & Eckert, 2013, on text characteristic “trade-offs”; Gervasi & Ambriola, 2003). That is, text characteristics at different linguistic levels may have conflicting impact on readers (their “targets”). For instance, an author may choose to write a text for second-grade students about a content-area topic, such as sound waves, requiring heavily laden vocabulary meanings that may make the text quite complex for young readers. But the words may also be technically challenging for word recognition. As an ensemble, difficult vocabulary meanings coupled with high decoding demand can magnify complexity exponentially. The author might consider ways of lessening the burden on the reader by employing other text-level characteristics, such as using a within-sentence syntactic pattern that is generally familiar to typically developing second-grade students or inserting parenthetical definitions after difficult word meanings, or at the discourse level, placing main ideas first in paragraphs. As another example, there is evidence that concreteness/abstractness, or imageability interacts with structural complexity and word familiarity to influence readers’ word recognition (e.g., Schwanenflugel & Akin, 1994). In short, constellations of co-occurring linguistic characteristics may contribute to variation in text complexity (Biber, 1988).

Measuring Text Complexity Quantitatively

Several established computerized systems address text complexity beyond the early grades through quantitative measurement. They are summarized here to provide context for the present study: readability formulae that are typically focused on word frequency, word length, and/or sentence length (e.g., Renaissance Learning, 2014; Klare, 1974; The REAP project, n.d.); conjoint measurement systems that relate students’ reading levels to text-complexity levels on the same scale, identifying collections of text characteristics (typically a small set such as word frequency and within-sentence syntax) that serve as “best predictors” of text complexity levels (e.g., the Lexile Framework for Reading [Stenner, Burdick, Sanford, & Burdick, 2006] and Degrees of Reading Power [DRP; Koslin, Zeno, & Koslin, 1987]); and natural language processing analyses involving multiple text characteristics (e.g., Coh-Metrix [Graesser et al., 2011; McNamara, Graesser, McCarthy, & Cai, 2014]; Reading Maturity Metric [Pearson Education, 2014]; and SourceRater [Sheehan, Kostin, Futagi, & Flor, 2010]). The systems may be differentiated in the following ways: (a) All measures except Coh-Metrix provide a single text-complexity quantitative judgment of texts’ complexity levels. Some do so using grade levels; others use their own leveling system. (b) Only Lexile and DRP measures are relational to readers, that is, they are originally based on individuals’ reading of the texts—except that the SourceRater measure uses an “inheritance principle” in which the original outcome variable used in the predictor equation was educators’/publishers’ assignment of text grade levels. Other measures examine text characteristics and then use a form of dimension reduction, such as principal components analysis to determine essential components of text complexity. (c) Coh-Metrix and SourceRater quantify the broadest number of text characteristics and include discourse-level text characteristics in their analyses.

Across the various systems, the most common text characteristics that are best predictors of text complexity are word familiarity, word length, sentence syntax, and/or sentence length. The SourceRater system involves eight dimensions—syntactic complexity, vocabulary difficulty, level of abstractness, referential cohesion, connective cohesion, degree of academic orientation, degree of narrative orientation, and paragraph structure. Coh-Metrix employs 53 text-characteristic measures reduced to five dimensions—narrativity, syntactic simplicity, word concreteness, referential cohesion, and deep cohesion. Importantly, none of the currently existing common metrics specifically provides explanation of what constitutes *early-grades* text complexity (cf. Graesser et al., 2011, and van der Sluis & van den Broek, 2010).

Summary

As the Common Core text-complexity standard is implemented in schools, educators and researchers alike need an empirically based understanding of text complexity for early-grades readers. Complexity theory provides a foundation for studying early-grades text complexity. Key principles of complex systems are that they involve a large number of mutually interacting parts; interplay among components can be locally, rather than globally, relevant; they often may be described by hierarchical organization; and they are purposive, that is engineered for particular purposes. A relational outlook on text complexity implies complexity of particular texts is relative to particular individuals, reading occasions, and developmental reading levels. However, theoretically, texts have an emergent “developmental” complexity such that they can be assigned a complexity level in relation to an entire continuum of complexity. Using an “optimality” concept in conjunction with what is known about critical cognitive factors for the early learning-to-read phase and prior findings about the importance of selected text characteristics during early reading, not only should many text characteristics at multiple linguistic levels be investigated, but interplay among text characteristics should be hypothesized. Few of the prior text-complexity measurement systems encompass discourse-level characteristics, few address text complexity as relational within either specific reading occasion or in the sense of student reading-ability development, none addresses the interplay or potential interactive nature of text characteristics, and importantly, none specifically addresses early-grades text complexity. In the present study, a relational frame is used to explore text characteristics that matter most for early-grades texts, and the potential interplay of text characteristics is naturally accounted for through use of a statistical modeling technique that is prevalent in many fields but novel to educational research, that is, random forest regression.

Method

Overview

Three hundred fifty primary-grades texts were selected and digitized. Twenty-two text-characteristics were identified at four linguistic levels. Multiple computerized variable operationalizations were created for each of the 22 text characteristics, totaling 238 variables. The variables were automated so that a computer could examine the digitized texts and produce text-complexity

measures for each operationalization. Analyses were conducted using a logical analytical progression typically used in machine-learning research (Mohri, Rostamizadeh, & Talwalkar, 2012). Three phases of analyses were: variable selection to find a subset of the most important text characteristics out of the 238 operationalizations; using 80% of the texts, “training” a random forest regression model (Breiman, 2001a) of the most important text characteristics associated with text-complexity level; and validating the model on a 20% “hold-out” set of texts. Follow-up analyses were done to explore the data structure.

Texts

Three hundred fifty texts (148,068 words in total) intended for kindergarten through second grade constituted the text base. An existing larger corpus of early-grades texts was made available for the study (MetaMetrics, n.d.-a), and maximum-variation purposive selection (Patton, 1990) was used to choose texts from the corpus. As well, 18 kindergarten through second-grade Common Core State Standards (NGA & CCSSO, 2010, Appendix B) exemplar texts (that were not present in the available corpus) were purchased. The goal of maximum-variation purposive selection was to ensure comprehensive representation of a wide variety of early-grades text types, text levels, and publishers that currently exist in U.S. early-grades classrooms. We chose 350 texts for two main reasons: (a) to include a sufficiently large number of texts that would adequately represent the domain and to ensure sound statistical analyses (following the suggested sample size in Heldsinger & Humphry, 2010) and (b) to include a manageable set of texts to accomplish teacher and student tasks needed for development of the text-complexity-level variable (described below in the section, “Text-Complexity Level”). All texts were reproduced in authentic form (including pictures) and digitized.

Six categories for commonly occurring early-grades text types for independent reading were determined: code-based (decodable, phonics), whole-word (texts that include many words that appear in early-grades texts with high frequency), trade books (books commonly sold for library, supplemental materials for classroom use, or private sale), leveled books (texts that are sequenced in difficulty level), texts of assessments, and other (e.g., label books). The first four text types had been previously identified in studies of classroom texts as reasonably comprehensive categories of early-grades texts intended for independent reading in primary-grade classrooms (Aukerman, 1984; Hiebert, 2011). The last two categories were included because texts appearing in assessments also commonly occur in early-grades classrooms, and texts of assessments may become even more prominent with the advent of the Common Core State Standards (NGA & CCSSO, 2010). Some commonly occurring early-grades texts, such as label books, do not fit well into the previous categories. The first four category labels are common terms used by educators and publishers (Mesmer, 2006).

It was not possible to consider proportional representation of types as they exist in United States classrooms because, to our knowledge, there is no direct evidence of the degree to which different categories of early-grades texts are present or used in U.S. classrooms, although at least one survey of U.S. primary-grades teachers suggested that use of the first four categories of texts is widespread (Mesmer, 2006). Consequently, we selected

“prototypes” to represent each category (Hiebert & Pearson, 2010), and, where series existed, texts were sampled from the range in the series. In reality, many early-grades texts fall into two or more of the category types (Mesmer, 2006). For example code-based texts are often “leveled.” However, for our purposes of ensuring wide representation of text types, each text was assigned to a single category. If a text was labeled “decodable” or “phonics” by the publisher, it was labeled “code-based.” If a publisher characterized a text as primarily attending to high-frequency words or sight words, it was labeled “whole word.” A text was labeled “trade book” if it was available in the trade market and not just in the school market, and it was not identified by the publisher as decodable, phonics, or high-frequency. A text was labeled “leveled” if the text was assigned a level (other than grade level) by the publisher and was not labeled “decodable,” “phonics,” or “high frequency.”

Text levels were determined by using publisher-designated grade, level, or age ranges. Texts were labeled: easy if they were designated kindergarten, kindergarten levels (as noted on publisher websites), or typical ages for kindergarten; moderately hard if designated first grade, first-grade levels, or first-grade ages; and hard if designated second grade, second-grade levels, or second-grade ages.

Thirty-two publishers were represented in the 350 texts, ranging from three to 15 different publishers for each of five of the six text types, with one publisher for the text-of-assessment type.

Text genre (narrative, informational, hybrid) was determined using a modification of Duke’s (2000) procedures. Two primary text characteristics were used to discern narrative, informational, and hybrid text—purpose and textual attributes. Narrative text was defined as follows (Duke, 2000; Rudrum, 2005): It is a series or sequence of events, with the intention or purpose to evoke an element of reader response. It tells a “story” and/or has characters, places events, and things that are familiar and is closely related to oral conversation. Informational was defined as text that conveys information about the natural or social world and is typically written by someone who is presumed to know the information to someone who is presumed to not know it (Duke, 2000). Textual attributes for narratives included for instance, events, actions with temporal or causal links, characters, dialogue. Textual attributes for informational texts included for example facts, timeless verb constructions, technical vocabulary, descriptions of attributes, definitions. A set of rules modified from Duke (2000) was devised for determining genre classification, using a decision tree process that began by determining the purpose of the book and then addressing attributes of the text. Interclassifier reliability between two individuals for 20% of the 350 books was .96.

Finally, the text corpus could be described as follows. Caution should be exercised when interpreting the following figures for the text categories—again, because the categories are not mutually exclusive. Rather, using the publisher designation in concert with the researcher-devised system described above for when a text could belong to two or more categories, 41% of the texts were leveled, 17% were code-based, 15% were trade books, 10% were whole-word, 9% were texts of tests, and 8% were other. Approximately 36% of the 350 texts were labeled easiest, 37% moderately hard, and 27% hardest. Sixty-six percent were labeled narrative, 24% informational, and 10% hybrid or other.

Variables

Text-complexity level. The outcome variable was early-reader text-complexity level measured using a continuous, developmental scale, with scores ranging from 0 to 100. An overview of the scale-building procedures is as follows. (Further details of the procedures are provided in the online supplemental materials.) Because text complexity was defined at the intersection of printed texts with students reading them for particular purposes and doing particular tasks, a multiple-perspective measure of text complexity was created using student responses during a reading task and teachers' ordering of texts according to complexity. In doing so, we represented students and teachers as readers, and teachers as important context for student reading instruction, as well as two different tasks in the final measure. Then the magnitude and strength of the association between the two logit scales (one from student responses and one from teachers' text ordering) was examined, and to arrive at a single scale, a linear equating linking procedure (Kolen & Brennan, 2004) was used to bring the student results onto a common scale with the teacher results. Finally, for ease of interpretability, the logit scale was linearly transformed to a 0 to 100 scale.

In a first substudy, through Rasch modeling (Bond & Fox, 2007) a text-complexity logit scale was created from the interface of 1,258 children from 10 U.S. states reading passages from a subset of the 350 texts and responding to a maze task (see Shin, Deno, & Espin, 2000, for task validity). Cronbach's alpha estimates of reliability for all test forms ranged from .85 to .96. Also, dimensionality assessments for text genre and for differential text ordering according to student ethnicity, gender, or free-reduced-lunch status suggested no evidence of measurement multidimensionality. After creation of the logit scale, each text in the subset was assigned a text-complexity level.

In a second substudy, also through Rasch modeling, a second text-complexity logit scale was created from 90 practicing primary-grades teachers' (from 33 states and 75 school districts) evaluations of texts' complexity. Teachers ordered random pairs of the 350 texts seen side by side on a computer screen. For each pair, teachers clicked on the text they thought was more complex. Determined by the separation index method (Wright & Stone, 1999), measurement reliability was .99. After creation of the logit scale, each of the 350 texts was assigned a text-complexity level.

Next, the correlation between the two logit scales ($N = 89$ texts) was .79 ($p < .01$), suggesting that the texts ordered on text complexity similarly whether teachers or students were involved. The relatively high correlation was also evidence of concurrent validity in that it suggested that the two logit scales were measuring the same construct. Consequently, a linking equating procedure was used to link the two logit scales (Kolen & Brennan, 2004). Finally, a linear transformation was done resulting in measures that could range from 0 to 100 on a text-complexity scale. That is, the 350 texts ordered by teachers could be assigned a measure from 0 to 100, and the texts read by students could be assigned a measure from 0 to 100.

Text characteristics and their variable operationalizations. Twenty-two text characteristics were identified at four linguistic levels—sounds in words, words, within-sentence syntax, and across-sentences or discourse level. Discourse-level characteristics captured repetition, redundancy, and patterning (of letters, words,

phrases, and/or sentences) that occurred in the texts. In an effort to capture a wide variety of ways of representing the text characteristics, multiple computerized variable operationalizations were created for many of the 22 text characteristics, totaling 238 variable operationalizations. The rationale for including as many variable operationalizations as possible was that different metrics may pinpoint different aspects of a text characteristic (Baca-Garcia et al., 2007). By including as many operationalizations as possible, the chances of capturing critical text characteristics for text complexity were increased.

Table 1 shows the 22 text characteristics according to linguistic level, along with definitions, the number of variable operationalizations for each, and selected examples of operationalizations and their possible score ranges and interpretations. A complete list and description of operationalizations is available in the online supplemental materials.

Operationalizations were accomplished using four logical approaches. First, several types of computational metrics were considered. In addition to traditional metrics, such as counts, mean, and percentage, six specialized computational linguistic techniques were used to produce other metrics. One specialized computational linguistic technique was distributional semantics (Landauer & Dumais, 1997), a method for quantifying semantic similarities between linguistic items. Three additional specialized computational linguistics techniques were: part-of-speech tagging (Collins, 2002); syntactic parsing (Sleator & Temperley, 1991); and a Levenshtein (1965) metric, which gauges the minimum number of substitutions, insertions, or deletions required to turn one linguistic unit (e.g., a written word) into another. Also, two unique metrics that specifically capture text characteristics in relation to student readers were applied to all of the sounds-in-words variables and most of the word-level variables—types- (unique words in a text) -as-test and words- (all words in a text) -as-test. Both metrics treat the text characteristic of interest as test items, while considering a potential student who might be reading the text to have a trait level for the characteristic of interest. Both represent an alternative way to measure central tendency for a distribution of values, and both are more impacted by outliers than an average. For instance, for a types-as-test operationalization for syllables (the text characteristic of interest) in a text, the unique words in the text are listed, and the number of syllables is counted in each word. Then one might hypothesize that a student has a "syllable-level reading ability" for reading the text. The unique words (types) form a test for measuring a student's ability to use syllables to read the text. Each unique word is given an item difficulty level that is the number of syllables in the word. A target level of hypothetical student performance is set (50%, 75%, 100% of the items predicted to be correct), and then using Rasch modeling (Bond & Fox, 2007) the metric determines what level of reader ability would be expected to attain the percentage that was set. The overall metric (derived from a mathematical formula) therefore summarizes a "syllable" level of complexity for the text.

A second logical approach was that discourse text characteristics were systematically treated as follows. The main focus of discourse-level variables was to capture linkages among words and meanings in text (e.g., cohesion), redundancy, and patterning that occur across a whole text or parts of text but more than just within sentences. For each discourse text characteristic, first, variable operationalizations were considered that would reflect a lexical

Table 1

Text Characteristics by Linguistic Level, Definition, Possible Score Range for Examples of Operationalizations, and Number of Operationalizations With Examples

Linguistic level	Text characteristic	Definition (source)	Possible score range for examples	Operationalization example (number of operationalizations)
Sounds in words	Number of phonemes in words	Smallest unit of sound. (The MRC Psycholinguistic Database provides phoneme values for words; Coltheart, 1981.)	1 (fewer phonemes in words, less complex) to less than 10 (more phonemes in words, more complex)	Mean number of phonemes for words in the text (14)
	Phonemic Levenshtein distance	The degree to which co-occurring phonemes exist across words. (Levenshtein Distance is a standard computer metric of string edit distance which gauges the minimum number of substitution, insertion, or deletion operations to turn one word into another. Measures phonemic similarity across words for the 20 closest words; Levenshtein, 1965; Yarkoni, Balota, & Yap, 2008; cf. Kruskal, 1999; Nerbonne & Heeringa, 2001; Sanders & Chinn, 2009.)	1 (few words in closest 20 share phonemes) to 3 (more words in closest 20 share phonemes)	Mean phonemic Levenshtein distance 20 with stop list 50 most frequent words (14)
	Mean internal phonemic predictability	The degree to which phoneme collocations occur given the totality of the phoneme collocations in the particular text. (Words are converted to phonemes using the Carnegie Mellon University, n.d., Pronouncing Dictionary.)	0 (fewer phoneme collocations are repeated in the text) to 1 (more phoneme collocations are repeated in the text)	Mean with text chunk size 125 (4)
Word structure	Decoding demand	The decoding demand of words in the text (slight modification of Menon & Hiebert's, 1999, decodability scale).	1 (less complex word structure) to 9 (most complex word structure)	Mean with stop list 50 most frequent words (22)
	Orthographic Levenshtein distance	See phonemic Levenshtein distance above. Orthographic Levenshtein distance measures orthographic similarity across words for the 20 closest words (Levenshtein, 1965; cf., Kruskal, 1999; Yarkoni, et al., 2008).	1 (fewer words in 20 share orthographic patterns) to 3 (more orthographic patterns shared)	Mean orthographic Levenshtein distance 20 with stop list 50 most frequent words (14)
	Number of syllables in words	Number of syllables in words. (The MRC Psycholinguistic Database provides syllable values for words; Coltheart, 1981.)	1 (few words with many syllables) to 8 (more words with more syllables)	Types as test with stop list 50 most frequent (ability at 75%) (18)
Mean internal orthographic predictability	Mean internal orthographic predictability	The degree to which letter collocations occur given the totality of the letter collocations in the particular text (researcher computer coded; cf. Solso, Barbuto, & Juel, 1979).	0 (fewer orthographic trigrams are repeated in the text) to 1 (more are repeated in the text)	Product of internal word values with chunk size 125 (4)
	Sight words	The most commonly occurring words in primary-grades texts (Dolch word list, n.d.; Fry Word List, 2012).	0 (less complex) to 100 (more complex)	Percentage of words in a text that are on the Dolch Preprimer list - (13)

Table 1 (continued)

Linguistic level	Text characteristic	Definition (source)	Possible score range for examples	Operationalization example (number of operationalizations)
Word meaning	Age of acquisition	Age at which a word's meaning is first known (Kuperman, Stadthagen-Gonzalez, & Brysbaert, 2012).	1 to 25 in our study (lower means more of the words are known by younger readers and a higher score means fewer are known by younger readers)	Age of acquisition types as test with stop list 50 most frequent words (ability at 50%) (13)
	Abstractness	Degree to which the text contains words that reference general or complex concepts such as "honesty" and cannot be seen or imaged (Paivio, Yuille, & Madigan, 1968; updated by Coltheart, 1981).	0 (less abstract, less complex) to 700 (more abstract, more complex)	Degree of abstractness types as test with stop list 50 most frequent words (ability at 50%) (20)
	Word rareness	The inverse of the frequency with which a word appears in running text in a corpus of 1.39 billion words from 93,000 kindergarten through university texts normalized to equate to the frequencies in the Carroll, Davies, & Richman (1971) frequency 5 million word list (MetaMetrics, n.d.-b).	0.10 (less rare, less complex) to 6 (more rare, more complex)	Word rareness types as test (ability at 90%) (14)
Syntax (within-sentence)	Sentence length	Number of characters, words, unique words, or phrases in a sentence (researcher computer coded).	1 (fewer characters, words, unique words, or phrases) and above 1 (more characters, words, unique words, or phrases)	Mean number of letters and spaces in sentences (6)
	Grammar	Link type, a linguistic convention that ties a word in a sentence to another word within the sentence. Differentiates between long sentences with many different syntactic relationships and long sentences with few syntactic relationships (Temperley et al., 2012; Sleator & Temperley, 1991). ^a	1 (fewer unique syntactic relationships, e.g., subject/object or noun-acting-as-adjective) to 29 (more unique syntactic relationships within sentences; a larger number can occur when the text has one or more very long sentences)	Mean number of unique link types in sentences (1)
Discourse (across sentences)	Family 1: Intersentential complexity			
	Linear edit distance	The degree of word, phrase, and letter pattern repetition across <i>adjacent</i> sentences. The number of single character replacements required to turn one sentence into the next one (Levenshtein, 1965).	0 (if all sentences are identical or there is only one sentence; lots of redundancy, less complex) to approximately 110 in our study (not much redundancy, more complex)	Mean linear edit distance (4)
	Linear word overlap	Degree to which unique words in a first sentence are repeated in a following sentence, comparing sentence pairs sequentially (researcher computer coded).	0 (no words are repeated in a following sentence) to 24.56 in our study (many words are repeated in a following sentence)	Mean linear word overlap with slice 125 (6)
Family 2: Lexical/syntactic diversity	Cohesion triggers	Words that indicate occurrence of cohesion in text. Five categories of cohesive devices between words in text work to hold a text together (cf. Halliday & Hasan, 1976; researcher devised beginning with words listed at Cohesion [linguistics], n.d.).	0 (no words on the cohesion trigger word list) to 39 in our study (many words on the cohesion trigger word list)	Percentage of words in text that are on the cohesion trigger word list (1)

(table continues)

Table 1 (continued)

Linguistic level	Text characteristic	Definition (source)	Possible score range for examples	Operationalization example (number of operationalizations)
	Type-token ratio	An indicator of word diversity, or the number of unique words in a text divided by the total number of words in a text (cf. Malvern, Richards, Chipere, & Durn, 2009).	0 (few unique words) to 1 (all words are unique)	Type-token ratio with chunk 125 (2)
	Family 3: Phrase diversity			
	Longest common string	Degree of word, phrase, and letter pattern repetition across <i>multiple</i> sentences. Captures couplets and triplets (Gusfield, 1997).	0 (a lot of overlap, a lot of redundancy, less complex) to 1 (not much overlap, more complex)	Mean Cartesian longest common string percentage with slice 125 (21)
	Edit distance	Number of single character additions, deletions, or replacements required to turn one string (or sentence) into another (Kruskal, 1999; Levenshtein, 1965).	0 (the same characters are repeated, high redundancy) to 127 in our study (very few characters are repeated, low redundancy)	Mean Cartesian edit distance with slice 125 (8)
	Cartesian word overlap	Degree to which unique words in a first sentence are repeated in a following sentence comparing all possible pairs in a 125 slice (researcher computer coded).	4 (unique words not repeated much in a following sentence) to 6 (unique words repeated more)	Percentage of mean Cartesian word overlap with slice 125 for part of speech (4)
	Family 4: Text density			
	Information load	Total information load in text. Denser texts have more information load, less redundancy, and are more complex. Also taps overlap of <i>groups</i> of co-occurring word repetition (researcher devised incorporating latent semantic analysis; Deerwester, Dumais, Furnas, Landauer, Harshman, 1990; Landauer & Dumais, 1997).	0 (low density, low information load, lots of novel co-occurring word-group repetition) to 1 (denser text, higher information load, not as much novel co-occurring word-group repetition)	Normalized percentage reduction of information load across sentences for 10 dimensions with slice 500 (12)
	Family 5: Noncompressibility			
	Compression ratio	The degree to which information in the text can be compressed. Novel text is less compressible (Burrows & Wheeler, 1994).	0 (more compressible, more redundancy, less complex) to 1 (less compressible)	Compression ratio with chunk 125 (2)

^a Definitions of all link types can be found at <http://www.link.cs.cmu.edu/link/dict/summarize-links.html>

emphasis or a syntactic (part of speech) emphasis. Second, whether an operationalization employed lexical or syntactic emphasis, operationalizations could also involve linear activity, that is adjacent sentences, or they could involve a Cartesian product over sentences (i.e., context beyond adjacent sentences), or they could address both types of activity. As an example, for the text characteristic, linear edit distance, the lexical-emphasis operationalization uses the words in two adjacent sentences, whereas a syntactical-emphasis operationalization uses parts of speech for replacement judgments. (Further detail is provided in the online supplemental materials.)

A third logical approach was to use existing databases and resources where possible to create variable operationalizations. The following databases were used. The MRC Psycholinguistic Database (Coltheart, 1981) "... is a machine usable dictionary containing 150,837 words with up to 26 linguistic and psycholinguistic attributes for each ..." (MRC Psycholinguistic Database, n.d., para. 1). Number of phonemes in words, number of

syllables in words, and indices of word abstractness were extracted from the MRC Psycholinguistic Database. The Carnegie Mellon University Pronouncing Dictionary (Carnegie Mellon University, n.d., "About the CMU dictionary," para. 1) "... is a machine-readable pronunciation dictionary for North American English that contains over 125,000 words and their transcriptions." It was used for variable operationalizations of the text characteristic, mean internal phonemic predictability. The Kuperman, Stadthagen-Gonzalez, and Brysbaert (2012) age-of-acquisition ratings for 30,000 English words was used for operationalizations of the age-of-acquisition text characteristic. The rating indicates the age at which a word's meaning is first known. Word frequencies for running text in a corpus of 1.39 billion words from 93,000 kindergarten through university texts (MetaMetrics, n.d.-b) normalized to link to Carroll, Davies, and Richman (1971) word frequencies, were used to create operationalizations for word rareness. The Link Grammar Parser (Sleator & Temperley, 1991; Temperley, Sleator, & Lafferty,

2012) was used for operationalizations of Grammar. The Parser “. . . is a syntactic parser of English, based on link grammar, an original theory of English syntax. Given a sentence, the system assigns to it a syntactic structure, which consists of a set of labeled links connecting pairs of words” (Temperley et al., 2012, para. 1).

Additional existing resources were as follows. The Menon and Hiebert (1999) decodability scale was slightly modified for operationalizations of the text characteristic, decoding demand. The scale provides numeric values for varying degrees of within-word structural complexity. The Dolch word list (n.d.) and the first 660 words on the Fry Word List (2012) lists were used in operationalizations of the text characteristic, sight words.

A fourth logical approach was to use techniques to control for factors that might be considered irrelevant to the measurement of specific text characteristics. One technique used for some operationalizations of sounds-in-words and word-level text characteristics was stop listing (Luhn, 1958), which is commonly used in natural language processing computations. Stop listing means deletion of the highest frequency words that tend to have low semantic value. However, because it is not known in advance whether deleting highly frequent words matters for examining text complexity, when stop listing was used for selected text characteristic operationalizations, the same text characteristics were also operationalized without stop listing.

Another technique was aimed at addressing possible impact of text length on a text-characteristic value. In general, longer discourse units can be related to increased complexity in part because inclusion of more material offers more opportunity for additional text characteristics or higher levels of individual text characteristics but also because each addition in a longer progression of discourse may require additional cognitive integration on the part of the reader (Merlini Barbaresi, 2003). Many text-characteristic operationalizations employed length control by using “slices” or “chunks” of text. When slices/chunks were employed, multiple slices/chunks were obtained from a text, covering the entire text, and then the final metrics were averaged over slices/chunks.

Analyses

Analyses were accomplished using a machine-learning logical analytical progression (Mohri et al., 2012). Random forest regression was used for statistical modeling. The analyses performed for the present study are among the first to appear in the educational research literature and therefore deserve some added attention and description here.

The statistical modeling approach. The interdisciplinary team of researchers who accomplished the present study worked from a statistical modeling approach that is not commonly used in educational research, but it is an approach that holds promise for some kinds of educational problems (Strobl, Malley, & Tutz, 2009). Two cultures of statistical modeling derive from diverse epistemological terrains in which different ways of knowing undergird different paradigms and procedures (Breiman, 2001b). A classical statistical modeling paradigm in educational research progresses in a top-down fashion. A theory is created detailing which constructs hypothetically matter in relation to some outcome(s) and how the constructs are related to one another. Consideration is given to how the constructs can be measured, a

relatively small set of “predictors” is selected, and the relationships are examined. Often a few interactions among predictors are hypothesized and represented in the statistical model. The resulting model is tested statistically through fit of the data to the originating model.

In another statistical culture, the one used in the present research, the counterculture to the predominant educational statistical paradigm, although theory can be involved initially (and was in our work), modeling works in a bottom-up fashion—starting with data (Breiman, 2001b). In the past years, multivariate data exploration methods have become increasingly popular in many scientific fields, including health sciences, biology, biostatistics, medicine, epidemiology, genetics, and most recently, psychology, and in machine-learning communities (Grömping, 2009; Strobl et al., 2009). “Machine learning” references construction, exploration, and study of algorithms and models that are “learned” or “trained” from data (Mitchell, 1997). Large amounts of data are processed, patterns are discovered, and predictor models are built. While some theoretical background is certainly helpful in discerning key constructs involved in a particular problem, there is no limit on the number of variables. Rather, all variables that can be imagined and measured are included as potential predictors. Sometimes, depending on modeling choice, any and all possible interactions among variables can be accounted for. The result is a model of the important predictors (and interactions) associated with the outcome. The “goodness” of the model is tested through its predictive capacity using a previously “unseen” set of data.

Random forest regression. The statistical modeling technique used in the present research was random forest regression—a nonparametric statistical analysis that involves an ensemble (or set) of regression trees (often referred to as CART—Classification and Regression Tree; Breiman, 2001a; Breiman, Friedman, Olshen, & Stone, 1984). Random forest regression overcomes limitations of a single regression tree and linear regression for particular circumstances such as when large numbers of variables are involved (Hastie, Tibshirani, & Friedman, 2009; Strobl et al., 2009). It is called an ensemble procedure because predictions from many decision trees are aggregated to produce a single prediction. Decision tree regression is based on the principle of recursive partitioning, where the feature space (defined by the predictor variable operationalizations) is recursively split into regions containing observations (in our case, texts) with similar response values. The predicted value for a text in a region is the mean of the response variables for all texts in that region. For example in our study, the many regressions produce regions or classes where texts have similar text characteristics in relation to their text-complexity levels. (For a detailed explanation of recursive partitioning, see Strobl et al., 2009.) The procedure is called *random forest* because each individual decision tree is “trained” using a different random bootstrap sample of the texts and because each split within each tree is created using a random subset of candidate variables (Grömping, 2009). (Bootstrapping is a process of repeated resampling of the data, with each sample randomly obtained with replacement from the original data set.) Ultimately, from the forest (ensemble) of trees, a single prediction can be made by calculating a mean of predictions output by the individual trees (Grömping, 2009).

Essentially, using the available data (in our case, the text-complexity level as outcome and 238 variable operationalizations

for each text as predictors), random forest regression builds a final model “from the ground up” by aggregating over many individually “trained” models. (To better understand random forest regression, and partly to better understand why it is potentially beneficial for analyzing text complexity, comparison to linear regression can be informative. A detailed comparison is provided in the online supplemental materials.)

Steps in analyses. Initially, an automated computer analysis was conducted for the 350 digitized texts and the 89 passages that students read, resulting in values for each text and passage for text-complexity level and for the 238 text-characteristic variable operationalizations. Then, four analytical phases were accomplished. (a) The first step in analysis was to set baseline performance. Eighty percent of the texts were randomly selected, and a three-pronged *training phase* was conducted using random forest regression. Three random forest regressions were conducted for: the 80% of the 350 texts that teachers ordered ($n = 279$; one text was discarded due to poor digitization), the 80% of the 89 student passages ($n = 71$), and the two sets of texts combined ($n = 350$). Each of the three random forest regressions yielded importance values for each of the 238 variables in relation to the text-complexity outcome variable. Model prediction capacity (correlation) and prediction error were calculated for each of the three models on “out-of-bag” samples (Grömping, 2009). (b) To determine whether a more parsimonious set of variables could predict text complexity as well as, or nearly as well as, the 238 variables, a two-stage *iterative variable-selection* procedure was used (Grömping, 2009). First, for each of the three models, the least important variable was removed from the model, random forest regression was rerun, and prediction error was recalculated. The process was repeated until model prediction error began to increase, resulting in a moderately sized set of predictors for each of the three models. Then the union of predictors in the three models was selected creating a moderately sized set of predictors. Second, in a next round of variable elimination, redundant operationalizations of text characteristics in the moderately sized set were identified, and the least important of the correlated redundant variables were trimmed out using a combination of strength of redundant operationalizations cut-point while maintaining model prediction capacity. (c) In a *validation* phase, the predictive capacity for the trimmed model was investigated, using texts not employed for the variable selection and “training” phases—a 20% hold-out set of texts. (d) Follow-up analyses were done to explore the data structure.

Results

Preliminary Random Forest Regression Decisions

The following decisions were made for conducting the random forest regressions using scikit-learn (Pedregosa et al., 2011): (a) At each node, the computer selected just one variable to make a split. (b) A constant predictor split point was used in each leaf. (c) Mean square error was used as the splitting objective to optimize in each node. (d) Randomness was injected into the trees using “bagging,” a method that allows all variables to be available for selection at a given node. During the training phase, “mtry” (the number of predictors available for selection) was set at 238. During the validation phase, “mtry” was set at three (or the square root of “p”

where “p” was nine predictors). The larger “mtry” was used when there was a moderate or large number of correlated predictors, because in the case of many predictors more power is concentrated in a relatively small subset of predictors. For variable selection, concentration of power is desirable, and as well, large mtry results in more stable variable selection because the most powerful variables tend to emerge repeatedly. (e) For variable selection, each random forest model was conducted with 100 trees. In the validation phase, random forest regressions were conducted with 500 trees. (f) The importance values were normalized random-permutation-based. (g) During training, out-of-bag model error (root-mean-square error [RMSE], for which error is normalized relative to the number of texts) was calculated as an estimate of generalizability error (Breiman, 2001a). During the validation phase, non-out-of-bag RMSE was calculated (Breiman, 2001a).

Phase 1. Training Phase Results: Baseline Model Performance

For the model using the 279 texts that teachers ordered and all 238 text-characteristic operationalizations, the mean correlation of text complexity as predicted from the model with the empirical text-complexity measures from 10 analytical runs of 100 trees each was .89, and the model error (RMSE) was 8.66. For the model using the 71 passages that students read and the 238 text-characteristic operationalizations, the mean correlation was .69, and the RMSE was 10.58. For the model combining the two sets of texts ($n = 350$), the mean correlation was .87, and the RMSE was 8.72. For each of the three models, predictive power was high, and error was low. (Importance values were computed for all 238 predictor variables in each of the three models, but given the large number of variables, only the final model variable importance values are reported in a following section.)

Phase 2. Trimmed Model and Final Operationalization Descriptives

First, Figure 1 shows that for two of the three models, as the least important operationalizations were dropped from the model, one by one, model correlation, that is, the predictive capacity, began to visibly drop for the teacher and combined models when approximately 25 variable operationalizations were left in the model. For the student model, it dropped with approximately 10 variables remaining. The union of the top 25 operationalizations in each of the three models was then selected, resulting in 45 predictor operationalizations. Then one model was created for the next step using the 45 predictor variable operationalizations.

Second, the first trim included redundant variable operationalizations for single text characteristics. To eliminate redundancies that were highly correlated, the intercorrelations of all 45 predictors were computed, using the combined data set. Then in the top of Figure 2 potential correlational thresholds are shown on the x-axis, and the y-axis shows what the model correlation would be if redundant variable operationalizations were removed using different magnitudes of threshold correlation as cut-points to delete redundant variables. Through visual inspection of the top graph, .70 was chosen as the correlational cut-point because it appeared that doing so would result in only very slight model correlation drop while removing a significant number of redundant predictors.

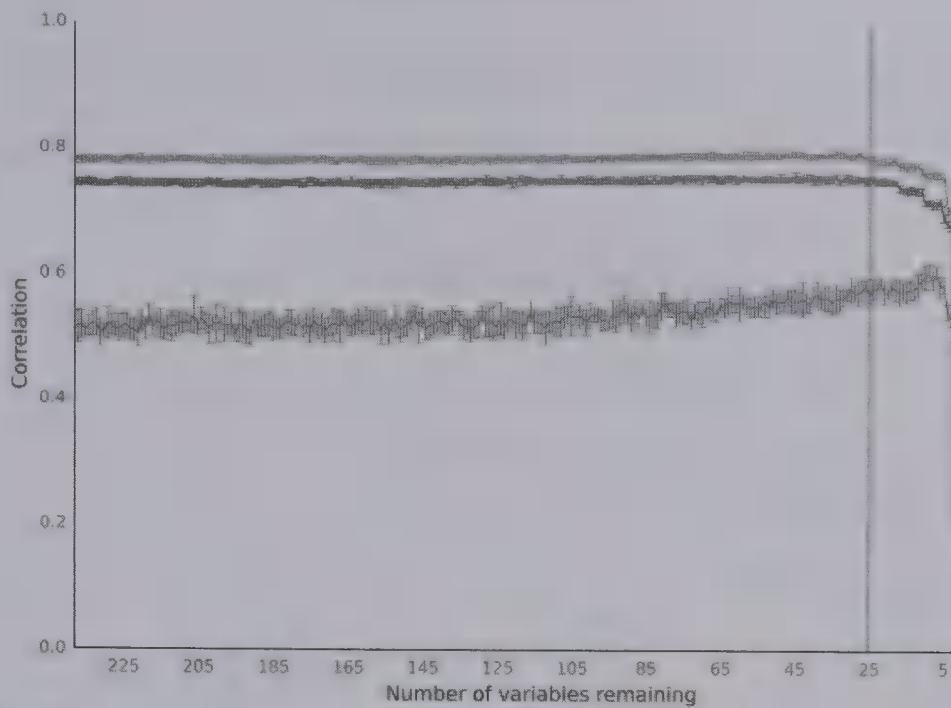


Figure 1. Correlation of predicted with empirical text-complexity in relation to least important variable deletion from each of three models. The top line represents correlational changes for teacher judgment, the middle line represents correlational changes for the combined teacher and student text-complexity assignments, and the bottom line represents correlational changes for the student text-complexity assignments. Also, out-of-bag correlation is used. At each point on the *x*-axis there is a central point that is the mean of the out-of-bag correlations from 10 independent random-forest runs. Surrounding the central point are error bars that represent the standard deviations from the 10 runs.

Then, as shown in the bottom graph in Figure 2, using the threshold cut-point of a .70 correlation, 11 variable operationalizations remained in the model. Among the 11, two sets of operationalizations were highly similar, and in each case, the least important of the two was dropped. In sum, the model trimming procedure resulted in a nine-predictor model, with the predictors noted here in parentheses by linguistic level: for word structure (decoding demand, number of syllables in words), for word meaning (age of acquisition, abstractness, and word rareness), and for sentence and discourse level (intersentential complexity, phrase diversity, text density/information load, and noncompressibility).

After variable selection, a final set of three random forest regression models was trained using only the nine variables (*mtry* = 3) with the teacher text-complexity assignments, the student assignments, and the two combined together. The resulting correlations (and RMSEs) for the teacher, student, and combined models were .89 (8.40), .71 (10.35), and .88 (8.59), respectively.

Phase 3: Model Validation

To validate the model, the hold-out set of 20% of books ($n = 71$) and 20% of the passages for student reading ($n = 19$) was combined. A final random forest regression (*mtry* = 3) was run with the nine selected variables as predictors and the empirical text-complexity variable from the combined (teacher and student) data as the outcome. The model was validated with a correlation of .85 and RMSE of 9.68. Figure 3 shows the generally tight relationship among the nine predictors and text complexity level. Variance explained by the model was 71.98%. Of note, the vali-

dation model error was similar to the combined data set model during training (8.72), suggesting minimal, if any, model overfit.

Variable Importance Values, Descriptives (Including Text Complexity), and Intercorrelations

Finally, after the validation phase, mean importance values were obtained from 10 final random forest regressions with 500 trees and *mtry* set at 3, using the 350 texts (Grömping, 2009). The variable importance values, mean, standard deviation, and range for the final nine variables along with mean, standard deviation, and range for the text-complexity variable, are shown in Table 2. The order of text-characteristic importance was: intersentential complexity (the linear edit distance operationalization; most important), text density/information load, phrase diversity (the longest common string operationalization), age of acquisition, number of syllables in words, abstractness, decoding demand, noncompressibility, and word rareness. Notably, three discourse-level characteristics appeared near the top of the importance order suggesting relative strength of discourse-level characteristics for predicting text complexity. Also included were word-structure and word-meaning text characteristics. While no variable that represented within-sentence text characteristic alone emerged, the discourse-level variables indirectly included facets of within-sentence characteristics—because to create measures across sentences, within-sentence characteristics had to be taken into account.

The text-characteristic variable operationalization means for the word structure variables (decoding demand and number of syllables

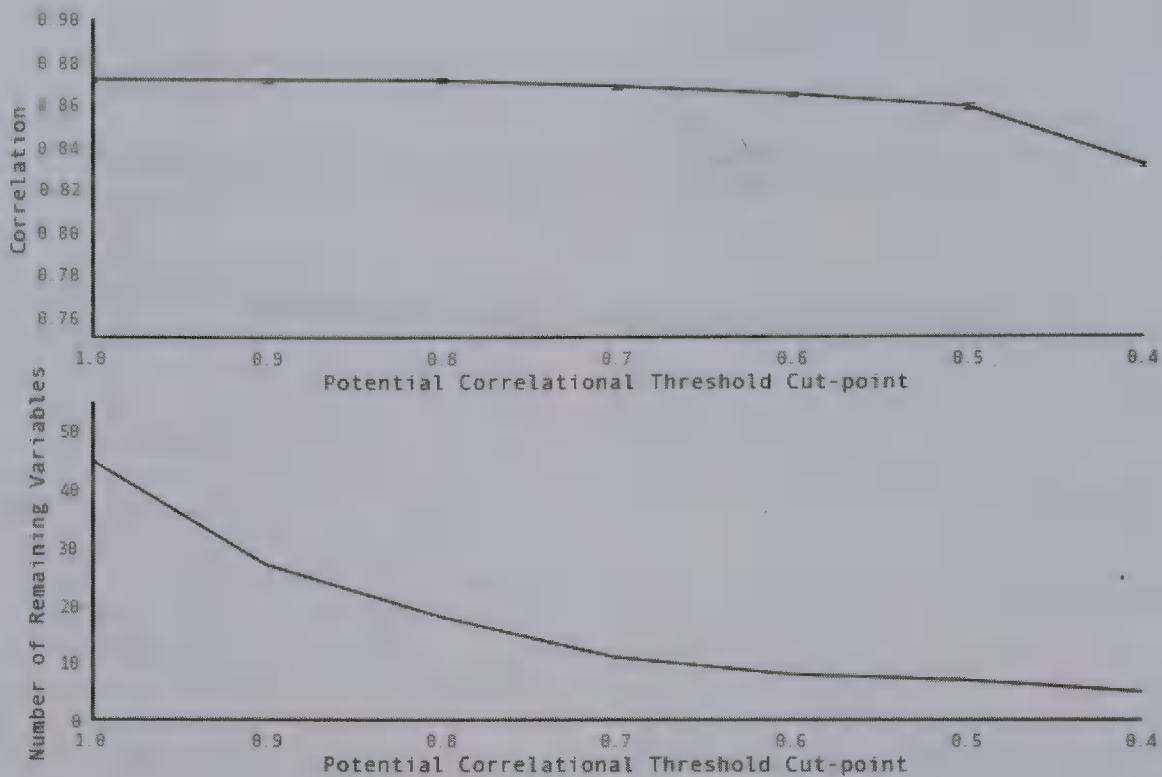


Figure 2. Trimming variables: relationship between potential correlational threshold cut-points (x-axis) with model correlation (y-axis; top figure) and the relationship between potential correlational threshold cut-points (x-axis) with number of remaining variables (y-axis). Correlation is the correlation of the predicted with the empirical text-complexity measure.

bles in a word) suggested that across the entire set of texts word structure was moderately challenging, though the range for decoding demand was wide—up to 7.91 (out of 9). (See Table 2 for summary statistics.) The means for the word meaning variable operationalizations (age of acquisition, abstractness, and word rareness) again suggested that on the whole, the abstractness of the words in the text was moderate (approximately at the middle of the possible range of scores), but as would be expected, word rareness

was minimal and age of acquisition tended to be low—though again, for all three variables, the standard deviations suggested a wide range of values. Means for the discourse level variable operationalizations suggested that the text corpus involved a fair amount of repetition, redundancy, and patterning in that means for three of the variables ranged from .55 (for noncompressibility, a compression ratio that could range from 0 to 1) to .80 (for phrase diversity, longest common string that could range from 0 to 1),

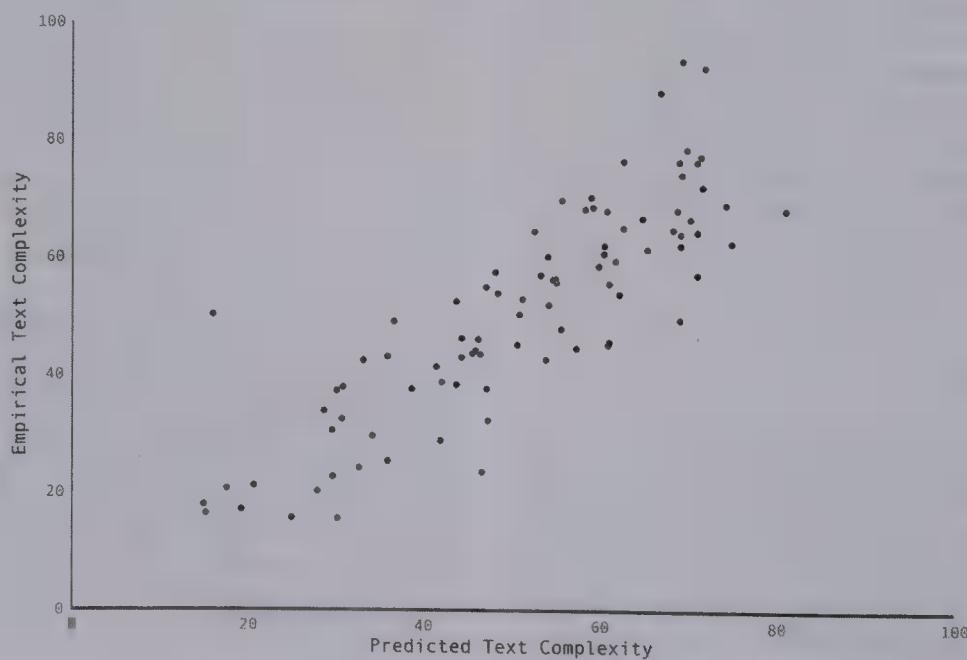


Figure 3. Scatterplot depicting the final model during validation.

Table 2
Importance Values for the Nine Text-Characteristics Variables and Descriptives for Text Characteristics and Text Complexity

Variable	Variable operationalization	<i>M</i> importance value (<i>SD</i>)	<i>M</i> (<i>SD</i>)	Range
Text complexity			50.10 (18.85)	0.33–100.00
Text characteristics				
Word structure				
Decoding demand (7)	Mean with stop list 50 most frequent words	.0164 (.0017)	5.32 (0.97)	2.00–7.91
Number of syllables in words (5)	Types as test with stop list 50 most frequent (ability at 75%)	.0633 (.0038)	1.42 (0.24)	0.00 ^a –2.42
Word meaning				
Age of acquisition (4)	Types as test with stop list 50 most frequent words (ability at 50%)	.0917 (.0073)	3.67 (0.52)	2.41–5.26
Abstractness (6)	Types as test with stop list 50 most frequent words (ability at 50%)	.0557 (.0040)	384.35 (63.11)	199.80–700.00
Word rareness (9)	Types as test (ability at 90%)	.0064 (.0004)	1.29 (0.29)	0.54–2.23
Discourse level				
Intersentential complexity (1)	Mean linear edit distance	.3487 (.0125)	31.04 (17.37)	0.00–109.88
Phrase diversity (3)	Mean Cartesian longest common string percentage with slice 125	.1782 (.0090)	0.80 (0.13)	0.31–1.00
Text density: Information load (2)	Normalized percent reduction of information load across sentences, 10 dimensions with slice 500	.2313 (.0116)	0.76 (0.10)	0.22–0.89
Noncompressibility (8)	Compression ratio with chunk 125	.0084 (.0006)	0.55 (0.11)	0.25–1.00

Note. Permutation accuracy importance values were used following Strobl, Malley, and Tutz (2009). Numbers in parentheses in the first column indicate rank order for importance value of descriptives for 350 texts, with values ranging from 1 (*most important*) to 9 (*least important*).

^a Zero scores occur when all the words in the text are on the stop list.

with intersentential complexity reflecting such features more modestly. In all four cases, nearly the complete range of values was represented in the corpus, suggesting a fair amount of variability on the discourse-level text characteristics. Finally, the full range of text-complexity values was witnessed, with a mean of 50.10.

The correlations in Figure 4 indicate moderately positive relationships of all nine variable operationalizations with text complexity, ranging from .35 to .73 with the exception of noncompressibility (.18, though significant). Next, on the whole, variable operationalizations within word structure, within word meaning (see the left-most triangle in Figure 4), and within discourse level (see the right-most triangle in Figure 4) were, on the whole, moderately correlated with each other, though in each of the three groups, there were one or two low correlations, suggesting that within linguistic level variable operationalizations tended to capture similar text characteristics. Also, on the whole, the cross-group correlations tended to be somewhat lower than within-group correlations, suggesting to some degree that each group of variables was measuring a unique set of characteristics (see the boxes in Figure 4). That is, correlations of decoding demand and number of syllables in words correlated with the three word meaning variable operationalizations from .06 to .54, all lower than .66, the correlation of decoding demand with number of syllables in words. The top right-most box shows a similar pattern. For the comparison of the word-meaning variable within-group correlations (the left-most triangle in Figure 4) versus the cross-group correlations of word meaning with discourse level variable operationalizations (the bottom box in Figure 4) again, on the whole, the within-group word-meaning correlations (.34 to .57), not including the low correlation of abstractness with word rareness (.05), tended to be similar to, or higher than, the cross-group comparison to the discourse-level

correlations (with the exception of the correlation of age of acquisition with intersentential complexity, .12 to .53).

Exploring the Data Structure and the Text-Characteristic Interplay

Several follow-up analyses (using all 350 texts and the teacher-based empirical text-complexity levels) were done to explore the data structure, the degree of text-characteristic variability in high versus low text-complexity levels, the interplay of text characteristics in relation to text complexity levels (decision trees and quintiles), and the interplay of text characteristics in relation to genre. The analyses were conducted using visualization methodology from CARTscans (a graphical tool that displays predicted values across multidimensional subspaces; Nason, Emerson, & LeBlanc, 2004), along with additional visualization techniques recommended by Cook and Swayne (2008) and by Cohen, Cohen, Aiken, and West (2003). A strong theme permeated findings—the interplay of text characteristics was an important factor for explaining text complexity.

The general structure of text characteristics in relation to text complexity. In a traditional approach, principal components analysis or factor analysis might be used to describe the data structure, but those techniques assume a linear relationship among variables. We hypothesized nonlinearity and used an unsupervised, nonlinear dimension-reduction technique—modified locally linear embedding analysis (Zhang & Wang, 2007). The technique accounts for the intrinsic geometric properties of each neighborhood of texts that share text-characteristic profiles. Essentially, in the analysis, the nine text characteristic operationalizations were re-expressed in a three-dimensional space by finding local planes of best fit for the neighborhood around each text (set at 15 neighbors;

	No. of Syllables in Words	Age of Acquisition	Abstractness	Word Rareness	Intersentential Complexity	Phrase Diversity	Text Density: Information Load	Non-Compressibility	Text Complexity
Decoding Demand	.66**	.49**	.17**	.30**	.45**	.31**	.37**	.16**	.47**
No. of Syllables in Words		.54**	.06	.37**	.51**	.42**	.34**	.18**	.51**
Age of Acquisition			.34**	.57**	.63**	.41**	.46**	.13*	.63**
Abstractness				.05	.34**	.37**	.53**	.12**	.49**
Word Rareness					.41**	.22**	.23**	.13*	.35**
Intersentential Complexity						.52**	.57**	.08	.73**
Phrase Diversity							.69**	.53**	.67**
Text Density: Information Load								.19**	.73**
Non-Compressibility									.18**

Figure 4. Correlations among the final nine text characteristics and text complexity. The top left-most box indicates the cross-group correlations for word-structure variable operationalization with the word-meaning variable operationalizations. The top right-most box indicates the cross-group correlations for word-structure variable operationalization with discourse-level variable operationalizations. The left-most triangle indicates within-group correlations for word-meaning variable operationalizations. The bottom box indicates the cross-group correlations for word-meaning variable operationalizations with discourse-level variable operationalizations. The right-most triangle indicates within-group correlations for discourse-level variable operationalizations.

* $p < .05$. ** $p < .01$.

Vanderplas & Connolly, 2009) and then stitching them together to describe the entire 350-text space. The planes of best fit need not share the same parameters across neighborhoods. Once the dimension-reduced text space was constructed, the text-complexity levels were noted in colors, warmer colors represent higher text-complexity levels, and cooler colors represent lower text-complexity levels. The result is shown in Figure 5. The three locally linear dimensions are not in themselves interpretable. Each is associated to varying degrees with the nine text characteristics. All 350 texts are represented as dots in the space. The main conclusion of the visual analysis was that there was a clear thread of text-characteristic relationships with each other and with text complexity that moved through the space, a thread that suggested an essentially unidimensional construct in measurement terms, *but* the text-characteristic relationships with text complexity were not globally linear. Instead, text-characteristic relationships interplayed differently in different local neighborhoods.

Degree of text-characteristic variability in high versus low text-complexity levels. To examine the extent to which text-characteristic variability was different according to text-complexity level, the nine text-characteristic variables were standardized as z -scores, and texts were split into high and low text-complexity groups using the following procedures (outlined in Cohen et al., 2003, and Green & Salkind, 2011). Centers for the high and low texts were determined at one standard deviation above and below the total text-set mean, respectively. Next, bands for high and low texts were created at plus and minus half of a standard deviation around the mean

of the center points, respectively, so as to filter out texts close to the mean (Cook & Swayne, 2008). Finally the split plots in Figure 6 were generated.

A main conclusion was that for most sets of relationships, there was more variability in lower text-complexity texts than in high ones. For the two word-structure relationships with text-complexity level, the decoding-demand levels for the low-complexity texts ranged widely, while most decoding-demand levels for high-complexity texts were tightly collected around the mean. For the two of the three word meaning characteristic operationalizations (age of acquisition and word rareness), the variability patterns were highly similar for low and high text-complexity texts, but for higher complexity, the word meaning values were shifted upward by approximately two standard deviations. On the other hand, for three of the four discourse-level variables (intersentential complexity, phrase diversity, and text density) there was little to no overlap in the two patterns, signaling a dramatic shift in the degree of repetition, redundancy, and patterning—less of it (higher values) in the higher complexity texts.

Also evident in the split plots are outlier texts. For instance, in the low text-complexity group for age of acquisition, there were some texts that had relatively high age-of-acquisition values, leading to the question of how a book with such high values on that text characteristic might receive a low value on text complexity. A general pattern appeared from examination of complete profiles of text characteristics for some randomly selected “outlier” texts.

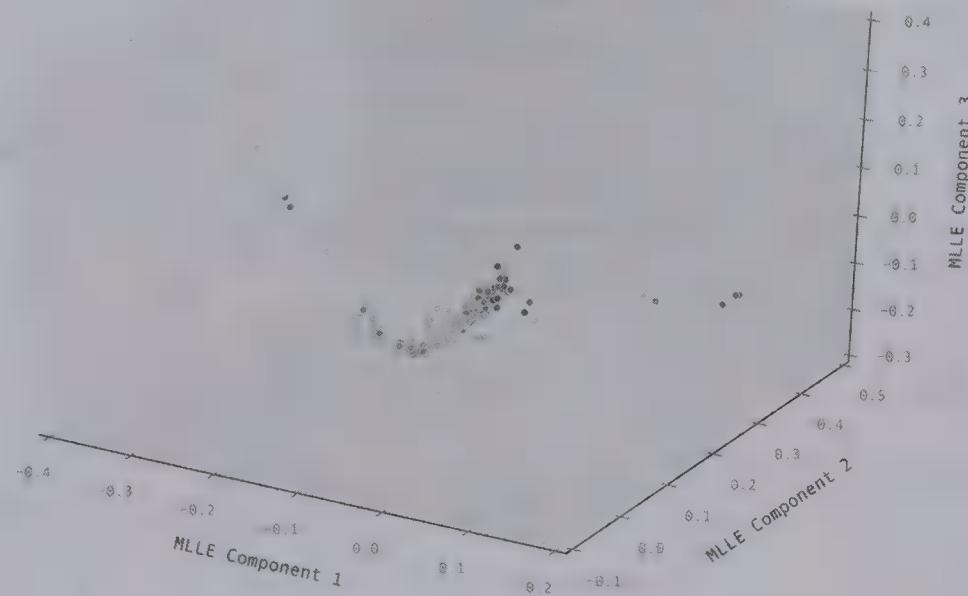


Figure 5. Three-dimensional scatterplot showing the data structure. Color represents text-complexity level, with red representing the highest text-complexity level, orange and then yellow the next highest, green and lighter blue moving lower, and blue the lowest. Each point is a text. MLLE = modified locally linear embedding.

Where extreme values were present in low text-complexity texts, generally, the high values tended to be compensated by low values on other text characteristics. For example, a text's relatively high value on a word structure or word meaning characteristic was modulated and supported by a high degree of repetition, sufficiently enough to effect a relatively low text-complexity level.

Interplay of text characteristics: Generalized interactions or regions of interactions? Two ways to explore the potential for text characteristics to function together in relation to text-complexity level were visualization of a single regression tree and contour plots (Nason et al., 2004). First, we created a single regression tree (see Figure 7) using standardized z -score values for the predictor variable operationalizations, with the tree grown to five levels of depth and restricting nodes to a minimum of 10 texts. The goal was to visualize the degree to which text characteristics might be conditioned on one another when predicting text complexity—not to determine which variables interacted with one another in the classic statistical sense. While information can be gleaned from exploring a single regression tree, generalization to early-reader texts at large is cautioned because of the possibility of single-tree overfit to a data set (Breiman 2001a).

Two main findings from examination of the decision tree were that the interplay of text characteristics mattered for text complexity and that microinteractions among text characteristics were regional rather than generally applicable to the whole body of text characteristics and text complexity. The tree depicts several localized interactions, or ways that text-complexity values may be predicted from combinations of certain text characteristics such that the impact of a text characteristic is conditioned by the value of one or more other text characteristics (two are noted in the dotted boxes in Figure 7). As an example, the far right side of the regression tree in Figure 7 depicts a localized asymmetrical interaction. Starting at the top of the regression tree in Figure 7, the computer algorithm

made the first split using intersentential complexity as the predictor that would result in the least error in predicting text complexity. To the right are texts that have intersentential complexity values higher than -0.3045 , that is, not much repetition, redundancy, or patterning. Moving farther to the right to Node B (which split the high intersentential complexity texts into even further subgroups of higher and lower intersentential complexity) and then Node C, the 109 texts at Node C have the least amount of repetition, redundancy, or patterning of the 350 texts. At Node C abstractness was selected as the predictor that conditioned intersentential complexity so as to achieve the smallest error in predicting text complexity. Notice that for 11 of the 109 texts, the ones with the lowest abstractness values, no further predictors were required to arrive at the final text complexity value with the smallest error. However, 98 of the 109 texts that had higher values on abstraction were further conditioned by noncompressibility and after that by age of acquisition. That is, the effect of abstractness is different for the two branches created by intersentential complexity.

Another interesting subtle finding reflecting the interplay of text characteristics that can be visualized from the regression tree is that sometimes slightly different combinations of text-characteristic conditioning can result in approximately the same text-complexity level. Notice for instance among the four left-most boxes just above the bottom row in the figure that two sets of texts have text complexity levels of 21.60 and 22.57, respectively. While both share similarly low intersentential complexity, for the left-most texts (21.60), conditioning intersentential complexity by the presence of higher word rareness values resulted in approximately the same text-complexity value as the right-most texts (22.57) where intersentential complexity was conditioned by lower values on noncompressibility.

A second way to explore potential interplay among variables was to visually examine contour plots (Nason et al., 2004).

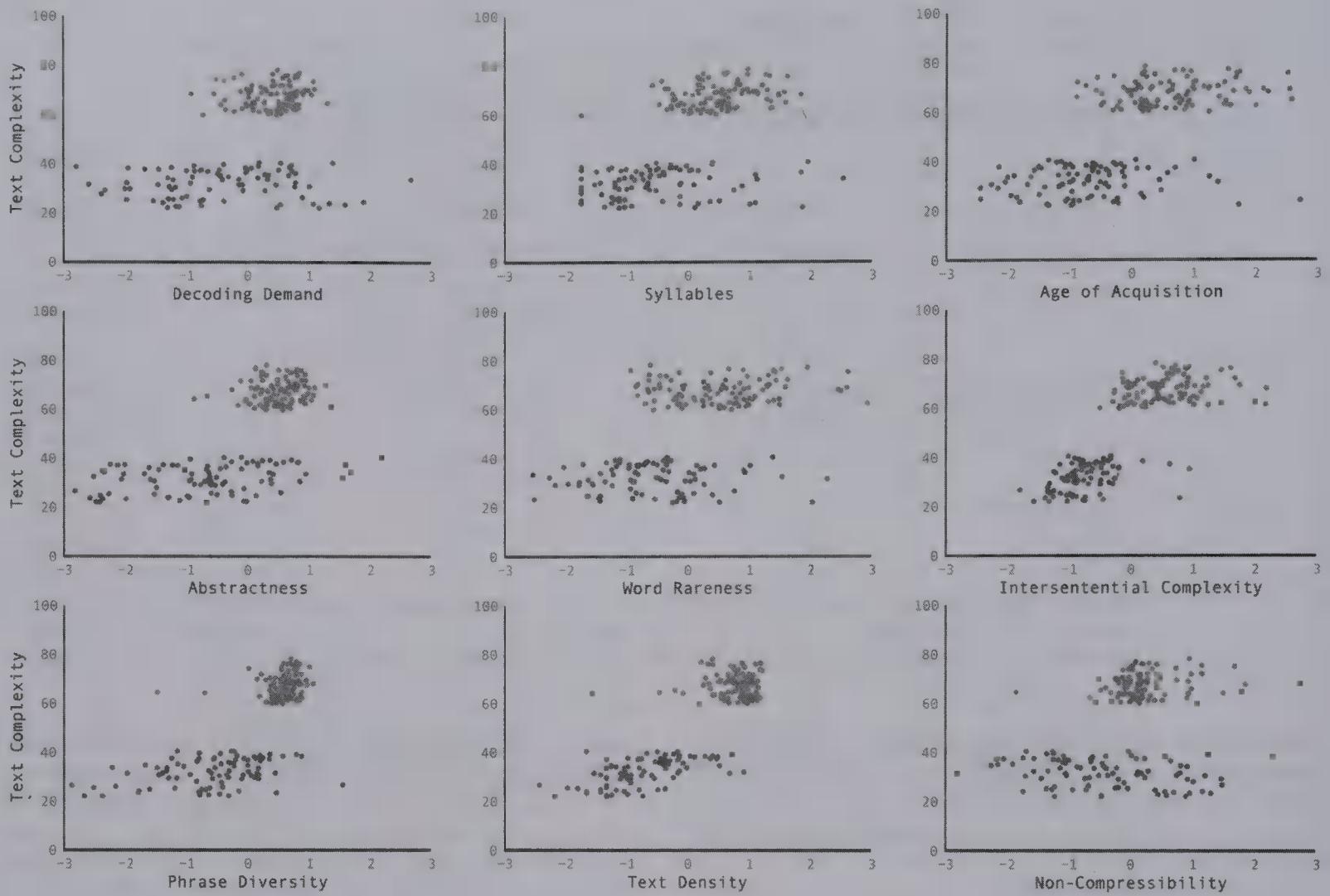


Figure 6. Split plots for individual text-characteristic variable relationships with low and high text-complexity levels. Top clusters (red/gray) are high text-complexity texts. Bottom clusters (blue/black) are low text-complexity texts. See the online article for the color version of this figure.

Several were created for selected combinations of text characteristics. A general finding was that there was interplay among the text characteristics in relation to text complexity. A limitation of contour plots is that a maximum of two predictors can be plotted. Figure 8 illustrates the interplay of age of acquisition with phrase diversity in relation to text-complexity level. The plot was generated from a random forest regression with just the two text-characteristic variable operationalizations and text-complexity level as the outcome, without controlling for the other seven text characteristics and with minimum node size of five. The main finding from the illustrative contour plot was that age of acquisition was conditioned by phrase diversity in relation to text complexity. Regions of texts are seen in the plot. The highest values on text complexity (red in the plot) occurred in texts that had high values on age of acquisition and high values on phrase diversity (low amounts of repetition, redundancy, or patterning). As well, texts with the lowest text-complexity values (dark blue) tended to have low values for age of acquisition and phrase diversity. However, some texts (e.g., light blue in the lower right quadrant) that had high values on age of acquisition had low text-complexity values when age of acquisition was moderated or conditioned by low values on phrase diversity, that is, when a fair amount of repetition,

redundancy, or patterning was present. The point is, again, there is interplay of text characteristics in relation to text-complexity level.

Text characteristic profile changes as text-complexity level increased. Another visualization method to understand text-characteristic collective patterning was to examine text characteristic profiles as text-complexity level increased (Cohen et al., 2003). The nine text characteristics were standardized as z -scores, texts were formed into quintile groups, and a graph was plotted using the within group means. As shown in Figure 9, first, the lowest quintile texts had a profile pattern that is markedly different from the other patterns. On average, the texts were characterized by less complex word structure (low decoding demand and relatively few syllables), relatively low-level vocabulary (younger age of acquisition, not very abstract words, and words that were not as rare as what appeared in more complex texts), coupled with, on the whole, highly redundant and repetitive texts (the exception is noncompressibility; recall that lower scores on the discourse level variables meant more redundancy and patterning). Moving up the graph, the next two quintile patterns were highly similar to one another, and the highest two quintile profiles were nearly flat with minor exceptions. In essence, text-characteristic profiles grad-

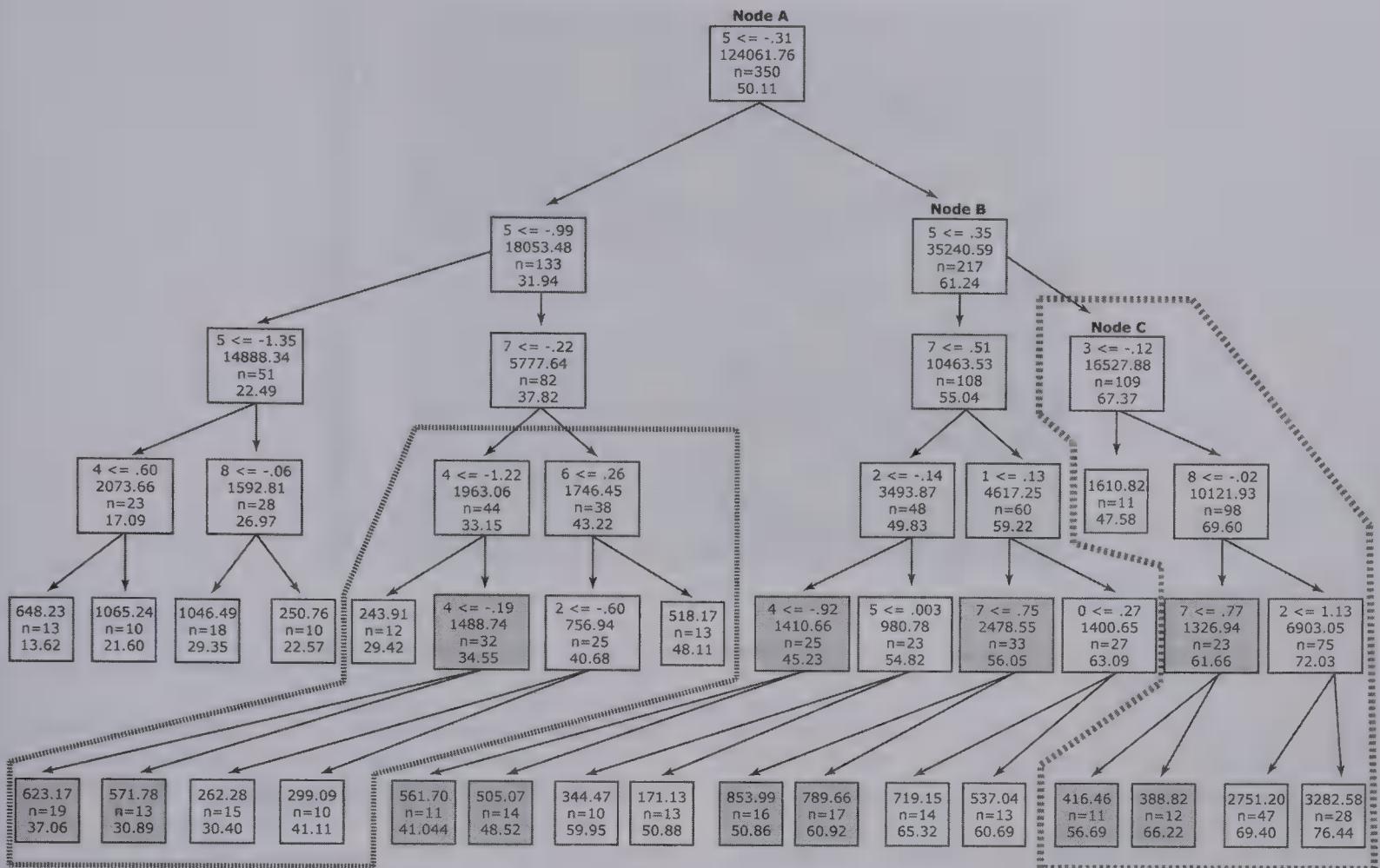


Figure 7. Single regression tree. The dotted lines surrounding selected boxes denote localized interactions among variable operationalizations. 0 = Decoding Demand; 1 = Syllables; 2 = Age of Acquisition; 3 = Abstractness; 4 = Word Rareness; 5 = Intersentential Complexity; 6 = Phrase Diversity; 7 = Text Density; 8 = Non-Compressibility.

ually changed as text complexity increased. Second, word structure became increasingly complex with each rising quintile. As well, on the whole, word meanings became harder and harder as text complexity increased. The exception was word rareness, which was similar in the bottom two quintiles. Also, on the whole, discourse-level redundancy and repetition decreased as text complexity increased (recall that higher discourse level averages reflected less redundancy and repetition). Noncompressibility was a minor exception in that although texts were consistently less compressible as text complexity increased, the changes were less dramatic than for other discourse-level variables or for word structure and word meaning characteristics. In short, on the whole, as text complexity increased, word structure and word meanings became harder, and texts displayed less and less redundancy, repetition, and patterning. Again, the interplay among the text characteristics was an important factor for text-complexity level.

Genre effects. Genre effects were analyzed using the same procedures as noted in the preceding section on “Degree of text-characteristic variability in high versus low text-complexity levels” (Cohen et al., 2003; Green & Salkind, 2011). Four groups of texts were created—narrative and informational texts that were high text complexity and narrative and informational texts that were low text complexity, and the

text-characteristic profile differences across genre, controlling for text-complexity level, were examined. Only texts identified as narrative or informational were included in the analysis because hybrid or other texts were rare. Text complexity means, standard deviations, ranges were comparable for the narrative and informational high text-complexity texts, and they were comparable for the two genres within low text-complexity texts: for high text-complexity narratives ($n = 64$), 67.16, 4.85, 59.86 to 78.19; for high text-complexity informational ($n = 24$), 67.39, 4.81, 60.34 to 77.02; for low text-complexity narratives ($n = 67$), 31.25, 5.56, 22.14 to 40.50; and for low text-complexity informational ($n = 17$), 32.88, 5.07, 24.11 to 40.43. Finally, the nine text characteristics were standardized as z -scores, and using the text-characteristic within-group means, the graph in Figure 10 was created to show the four text groups’ text-characteristic profiles.

In general, as would be expected, controlling for text-complexity level, the genres within text-complexity level had slightly different text-characteristic profiles. For high text-complexity narrative texts, on average, abstractness, intersentential complexity, phrase diversity, and text density tended to have higher levels than the other text characteristics. On the other hand, for high text-complexity informational texts, only age of acquisition, on average, tended to rise above the other text-

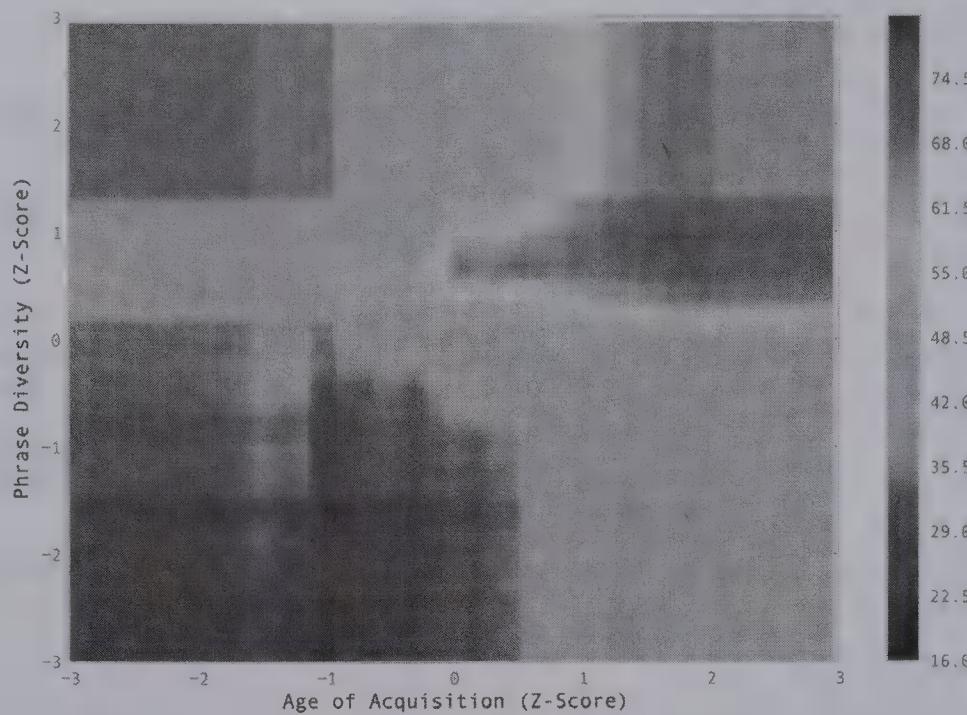


Figure 8. Contour plot of age of acquisition, phrase diversity, and text complexity.

characteristic levels, and, on average, noncompressibility tended to dip below all other text characteristic levels. Notably, several text characteristics were at approximately the same levels in the two genres. The most divergent characteristics across high-text-complexity text genres were age of acquisition (higher for informational texts) and word rareness (also higher for informational texts).

For low text-complexity narrative texts, on average, text-characteristic levels were approximately similar, with the exception of noncompressibility, which is, surprisingly, much higher than the others. For low text-complexity informational texts, on average, decoding demand, syllables, and word rareness tend to be

higher than the other informational text characteristics. Notably, several text characteristic levels were similar across the two low-text-complexity genres. The most divergent were decoding demand, syllables, word rareness—all higher measures for informational texts—and noncompressibility, which was higher for narratives. Again, another example of text-characteristic interplay was witnessed. When word structure and word meanings were relatively difficult (as for informational texts compared to narratives), more repetition and patterning at the discourse level (realized by relatively low scores) likely modulated the impact of the difficult words to bring the overall text complexity to a relatively low level.

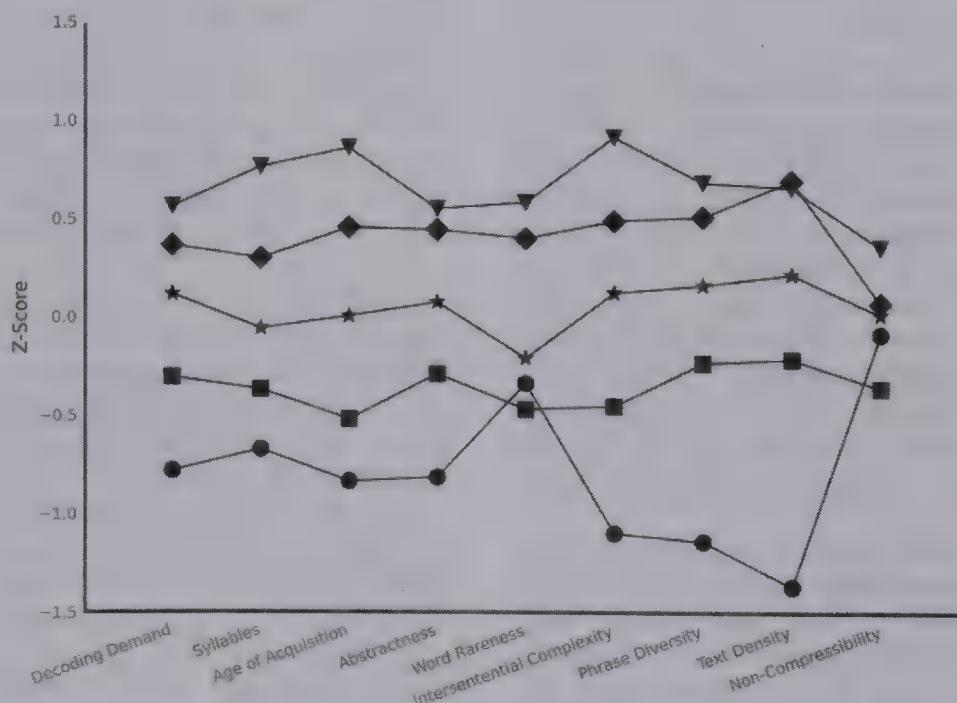


Figure 9. Text-characteristic profiles by text-complexity quintile group.

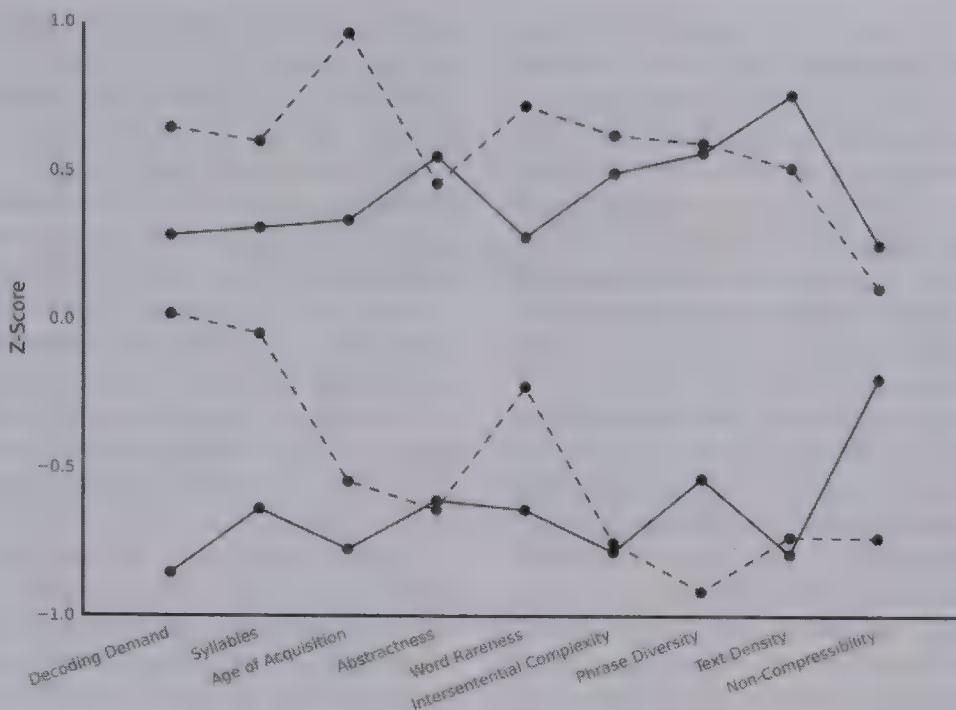


Figure 10. Text-characteristic profiles according to text-complexity level and genre. The top two lines represent high text-complexity levels. The bottom two lines represent low text-complexity levels. Solid lines represent narrative texts, and dotted lines represent informational text.

Conclusions and Discussion

Conclusions

Nine text characteristics were most important for early-grades text complexity: word structure—decoding demand and number of syllables in words; word meaning—age of acquisition, abstractness, and word rareness; and sentence and discourse level—intersentential complexity (the linear edit distance operationalization), phrase diversity (the longest common string operationalization), text density/information load, and noncompressibility. The nine-characteristic model predicted text complexity very well, in fact, nearly as well as the more complicated model with all 238 text-characteristic operationalizations. Notably, the three most important text characteristics were at the sentence and discourse level—intersentential complexity, text density/information load, and phrase diversity. Additionally, interplay among text characteristics was important to explanation of text complexity. While a clear thread of the relationship of the nine text characteristics with text complexity was evident, the relationship was not globally linear. Instead, text-characteristic relationships interplayed differentially in local neighborhoods of similar texts.

Discussion

To our knowledge, the present study is the first to reveal important text characteristics for early-grades text complexity through empirical investigation. The results support the contention that early-grades texts can be considered complex systems consisting of characteristics at multiple linguistic levels that variously interplay to impact text complexity. Further the nine most-important text characteristics revealed in the present study map to some of the well-researched critical features of young children's

early reading development. The early-grades developmental phase is often characterized as “cracking the code,” which has led some educators to believe the work of early reading is primarily about, or even all about, phonological awareness and word-related factors. Interestingly, phonemic measures did not surface among the most important text characteristics for text complexity. The importance of phonological awareness for progress in early reading is indisputable. Possibly the measures in the current study did not sufficiently reflect the domain of key phonological knowledge required of students.

As for the centrality of word structures in “cracking the code,” it was not surprising to find that word decoding and number of syllables were among the top-most important for predicting text complexity. As well, factors involved in word meanings, specifically age of acquisition of words, abstractness, and word rareness, were important. The findings are consistent with prior suggestions that lower text complexity might be achieved in part through inclusion of easier and more familiar vocabulary (e.g., Hiebert & Fisher, 2007).

At the same time, aspects of the findings in the present study shed additional light on the distinctiveness of early-grades text complexity compared to upper-grades text complexity. While traditional measures of within-sentence syntax (such as sentence length or various grammatical indices) were not among the nine most important text characteristics, some of the discourse-level metrics captured within-sentence complexity while also measuring text characteristics beyond the sentence level. For instance, while the intersentential complexity metric, linear edit distance, addressed the degree of word, phrase, and letter repetition across adjacent sentences, it was also impacted by overall sentence length irrespective of patterning and repetition. That is, linear edit distance captured both within and across-sentence characteristics. Consequently, within-sentence features were necessarily included.

Still, it is worth noting that traditional within-sentence indicators such as sentence-level syntax or sentence length itself were not among the critical metrics for early-grades text complexity. One possible reason is that although within-sentence indicators tend to be highly associated with complexity for texts beyond second grade, many early-grades texts that have long sentences tend to have long sentences that are marked by repetition of words or phrases. The repetition of words or phrases in early-grades texts may reduce the challenge posed by long sentences and render within-sentence indicators, such as length, less effective for estimating early-grades text complexity.

One of the most striking findings was the emergence of discourse-level text characteristics that primarily captured repetition, redundancy, and patterning in texts. The finding was striking because it is often *not* discussed in the context of “code cracking.” Educators and researchers tend to focus on word-level text characteristics as almost singularly critical for early reading, and the role of how texts are structured to facilitate ease of early-reading progress is often overlooked. Indeed, even one of the most commonly used text-leveling systems, the Fountas and Pinnell (1996, 2012) system, does not directly include attention to repetition and redundancy, though they do address text structure and genre in general. As noted earlier, few prior text-analysis systems for the upper grades include analysis of discourse-level characteristics—although those systems were not intended for early-grades texts. However, at least one or two of the discourse-level characteristics (intersentential complexity and phrase diversity) in the present study are reminiscent of cohesion operationalizations in the Coh-Metrix (Graesser et al., 2011) system. While some evidence exists that above second-grade level, models of text complexity that include discourse-level indicators do not outperform those that do *not* include them (Nelson, Perfetti, Liben, & Liben, 2011), our findings suggest that attention to discourse-level characteristics at the early grades is crucial (cf. Hiebert & Pearson, 2010, who suggested that current text-complexity systems may need adjustments for early-grades texts). Indeed, the functions of repetition and redundancy in discourse have received increasing attention on the part of linguists in the past few years, and repetition/redundancy is considered by some to be an essential feature of language use (Bazzanella, 2011).

Unearthing the presence of locally embedded differential interplay of text characteristics and witnessing examples of that interplay are novel contributions to the literature. The finding was intriguing in that to the mature eye, early-grades texts appear to be “simple.” But experienced readers often have long forgotten the challenges of learning to read in the early phases, and to more expert readers, as Prince (1997) and others (e.g., Bazzanella, 2011) have pointed out, “. . . the really interesting complexities of language work so smoothly that they become transparent” (Prince, 1997, p. 117).

The finding of locally embedded text-characteristic interplay was also supportive of prior linguists’ and complexity theorists’ understandings that in complex environments, subsystems (in the present study, sublinguistic systems) often “cooperate” to balance efficiency and effectiveness. In the case of early-grades texts, subsystems “cooperate” to balance young children’s ease of learning to read with the requirements for depth of processing (Bar-Yam, 1997; Juola, 2003; Merlini Barbaresi, 2003). However, while the presence of regional interactions among text characteristics

could be witnessed, as for example, in the single decision tree and the contour plot, explaining or describing them with simple generalizations was difficult because of the number of characteristics involved and the variation in coexisting characteristics across witnessed incidents of interactions.

Although local interplay was a chief characteristic of early-grades text complexity, some general trends described features of the early-grades texts in the aggregate. One general trend was that, on the whole, as text-complexity level increased, word structure and word meaning text characteristics became more complicated or harder (as would be expected), while texts displayed less and less redundancy, repetition, and patterning. That is, linguistic levels interplayed such that text characteristics tended to coalesce in one way for less complex texts and in another way for more complex texts.

Another general trend was for high-complexity informational texts to have somewhat higher age-of-acquisition and word rareness measures compared to narrative texts. On the other hand, low-complexity informational texts tended to have somewhat higher decoding demand, more syllables, and rarer words than narratives, but narratives were less compressible. For both high- and low-complexity texts, interestingly, discourse-level text characteristics were fairly similar across the two genres with informational texts having slightly lower discourse-level values, indicating more repetition, redundancy, or patterning. The result again supports the interplay of variables in that the presence of more difficult words was compensated by increased scaffolding in the form of repetition or patterning. The difference should be considered with caution, as a relatively small number of books constituted the genre analysis. Rather than assuming the result is generalizable, it is more appropriate to consider it sufficiently provoking to warrant further analysis in future studies.

However, taken at face value, the genre result is consistent with logical expectations. In general, at the early-grades levels, informational texts might tend to have more difficult vocabulary than narratives, and at the lowest text-complexity levels, it would be challenging to lower decoding demand for content-laden material. It is worth noting that when using random forest regression with the nine-characteristic text-complexity model, random forest regression easily accounts for any localized or general text characteristic collections that might be related to genre.

The promise of random forest regression and machine-learning research methods. The successful use of random forest regression for modeling text complexity in early-grades texts demonstrates the potential for the random forest regression advantage when addressing a high-dimensional educational problem. In the case of early-grades text complexity, a modeling technique such as linear regression may not satisfactorily allow for investigations employing either the large number of variables required for text analysis or the potentially huge number of complex text-characteristic interactions that likely permeate early-grades texts. It is important to note, however, that we did not accomplish a comparison of results from a theorized linear regression model and a random forest model, and consequently our statement here about the possible random forest regression advantage is hypothetical. At the same time, it is difficult to imagine how such a comparison could be tested—because there is no way to tap a priori localized interactions among text characteristics in traditional linear regression.

As well, random forest can be a more robust model than some other traditional modeling techniques in that it accounts for exceptional cases. To comprehensively study early-grades texts, where many different types of text exist, it is important to include even those texts that might traditionally be considered “outliers,” that is, texts that might have text-characteristic configurations that fall in the long tails of early-grades text distributions. For instance, label books do not contain connected text, but instead one word is shown beside a picture. In a traditional analysis, such books might be considered outliers because they have text characteristics that are quite different from a majority of texts. However, label books are commonly used in early-grades classrooms, and any study of text complexity should take them into consideration. As well, random forest regression automatically handles conditionality that can occur in ensembles of text characteristics, and as such it brings the tails of distributions “into the fold.”

Finally random forest regression can take advantage of a weak predictor by using it only when it is needed. In the present study, noncompressibility might be considered a weak predictor in that it was not highly correlated with other characteristics (except for phrase diversity) or with text complexity. However, noncompressibility tended to locate repetition, redundancy, and patterning where the other three discourse-level characteristics did not locate it. Such texts were rare in the present study, but on those rare occasions, there was important value in the noncompressibility measure.

High-dimensional problems are common in educational arenas in cases where large numbers of variables are at play and large amounts of data are generated, and random forest regression is a statistical modeling technique that could innovate the repertoire of educational statistical modeling. Where pressing educational problems involve large numbers of variables and/or potentially large numbers of interactions among variables, random forest regression could provide uniquely satisfying solutions (Baca-Garcia et al., 2007).

The machine-learning techniques used in the present study uniquely revealed early-grades text complexity. While prior text-complexity systems existed, theorization about text complexity, especially early-grades text complexity, was limited (Mesmer et al., 2012), and debates about construct coverage in the existing measurement systems proliferated (e.g., Sheehan et al., 2010). As a consequence, employing a wide array of possible operationalizations of text characteristics, each of which might capture a nuanced sense of any text characteristic, was important, as was the use of a logical investigative progression to narrow the most important characteristics. That is, through machine learning techniques, the data could “speak,” and a text-complexity model could be constructed from the data themselves (Wasserman, 2014).

Further, the interactive, dynamic graphics used to explore data structure are common in machine-learning communities, but not as common in educational research. While no statistical significance was attached to the visualization techniques, they tended to be very useful in understanding functional relationships among text characteristics and text complexity.

Limitations of the study. The following limitations of the study should be considered as context for interpreting the findings. First, although random forest provided many advantages for the study of early-grades text complexity, the resulting functional shape of the data was interpretable only to a certain degree. That

is, the complexity of text-characteristic interactions was acknowledged, but it could not be described in simple ways or with a parsimonious set of rules. Whether lack of a final specified statement detailing local interactions is a failure or a limitation is debatable. For those who embrace complexity theory, tensions between chaos and parsimony, between complexity and simplicity are natural—they exist in the natural world, and attempts to over-specify distort reality.

Second, text selection for study was extremely important. The population of classroom texts should be broadly represented. While every attempt was made to accomplish broad representation, the texts selected for the study may set boundaries on the generalizability of findings, and readers of the study should draw their own conclusions about the text representation.

A third limitation is that a traditionalist statistician working in the fields of psychology or education might consider the process of trimming variables awkward or imprecise. Lacking statistical estimation of variable “significance,” logical analysis was necessary. Some may question the reliability of the logical analysis. Certainly, when such methodology is used, it is critical that detailed description is provided so that readers may glean whether conclusions are warranted.

A fourth possible limitation is that because pictures could not be analyzed digitally, the role of pictures in early-grades text complexity was not directly assessed. However, pictures were indirectly involved in that they were present in both the teacher and student substudies for creation of the text-complexity metric.

Implications for practice. One major practical implication of the present results is that educators should consider discourse-level text characteristics in early-grade readers perhaps more than is the current case. Some researchers and teacher educators advocate that educators should account for text “organization” (e.g., Shanahan, Fisher, & Frey, 2012), or in the case of Coh-Metrix, discourse-level features such as cohesion (Graesser et al., 2011), when assigning texts to students. Given that “code-cracking” is prevalent during the early-grades, it is likely that in everyday classroom instruction, word-level characteristics are favored, and discourse-level text characteristics may be given short shrift. Instead, attention to discourse-level features such as repetition, redundancy, and patterning would appear to be in order.

As well, few teacher educators or researchers espouse the significance of the interplay among text characteristics for text complexity in general, even above the early grades. While the important text characteristics often, if not typically, make unique contributions to text complexity, in many texts, their interplay is equally important, if not more important. Consequently, it is critical that, when selecting texts for young children, educators consider ways in which characteristics can modulate one another’s challenges. For example, presence of repetition, redundancy, and patterning can ease reading progress for children when texts have somewhat challenging word structures and/or word meanings. In light of evidence that present-day core-reading programs tend to have somewhat difficult vocabulary (Foorman et al., 2004), teachers might particularly observe degrees of repetition and patterning in core readers and provide additional instructional support for students as needed.

The finding of more variability in lower text-complexity texts than in higher ones was interesting in that some might anticipate the opposite—less variability (more control over) the char-

acteristics for students who are just beginning to learn to read, with more variability (less control over) characteristics as students advance their reading ability. Educators might need to consider the lowest level texts especially carefully when choosing texts for students' independent reading versus for instructional settings where teachers can provide more support.

Finally, publishers of early-grades texts should account for multiple text characteristics when creating and/or leveling early-grades texts. Some current-day leveling systems that are commonly used by publishers and/or classroom teachers, such as Fountas and Pinnell's (2012) system, do take into account text characteristics at multiple linguistic levels, but many publishers rely solely on measurement of word frequency and sentence length. While the latter two factors can be useful for many reasons, creation of optimal texts that ease young students' reading growth and use of optimal leveling systems likely requires consideration of a wider gamut of early-grades text characteristics.

Implications for future research. The present findings lend credence to a complexity theory of early-grades texts. One challenge for future research is further exploration of potential classes of early-grades texts where, within class, selected ensembles of characteristics condition one another in similar ways. If such classes of texts are identifiable, through professional development sessions, educators might come to a fuller understanding of the importance of selecting texts with certain characteristics to enhance particular cognitions as students begin to learn to read.

The results of the present work suggest that a tool, an automated analyzer, could be created from the final nine-variable predictor model using random forest regression. The development of such a tool could be potentially useful to researchers who are interested in evaluating existing reading materials or to guide the development of new materials.

Finally, the present text-complexity model of text characteristics might also be used in intervention efforts. Texts could be theoretically configured as "best texts to facilitate young children's reading progress." Then in a controlled comparison-group intervention design, children's reading progress could be examined when reading instruction occurs with such texts compared to other classes of texts that exist widely in current-day classrooms.

References

ACT. (2006). *Reading between the lines: What the ACT reveals about college readiness in reading*. Iowa City, IA: Author.

Adams, M. (1990). *Beginning to read: Thinking and learning about print*. Cambridge, MA: MIT Press.

Albert, R., & Barabási, A.-L. (2002). Statistical mechanics of complex networks. *Reviews of Modern Physics*, 74, 47–97. doi:10.1103/RevModPhys.74.47

Aukerman, R. C. (1984). *Approaches to beginning reading* (2nd ed.). New York, NY: Wiley.

Baca-Garcia, E., Perez-Rodriguez, M. M., Saiz-Gonzalez, D., Basurte-Villamor, I., Saiz-Ruiz, J., Leiva-Murillo, J. M., . . . de Leon, J. (2007). Variables associated with familial suicide attempts in a sample of suicide attempters. *Progress in Neuro-Psychopharmacology & Biological Psychiatry*, 31, 1312–1316. doi:10.1016/j.pnpbp.2007.05.019

Bar-Yam, Y. (1997). *Dynamics of complex systems*. Reading, MA: Addison Wesley.

Bazzanella, C. (2011). Redundancy, repetition, and intensity in discourse. *Language Sciences*, 33, 243–254. doi:10.1016/j.langsci.2010.10.002

Biber, D. (1988). *Variation across speech and writing*. Cambridge, England: Cambridge University Press. doi:10.1017/CBO9780511621024

Bond, T. G., & Fox, C. M. (2007). *Fundamental measurement in the human sciences* (2nd ed.). Mahwah, NJ: Erlbaum.

Bowers, P. G., & Wolf, M. (1993). Theoretical links among naming speed, precise timing mechanisms and orthographic skill in dyslexia. *Reading and Writing*, 5, 69–85. doi:10.1007/BF01026919

Breiman, L. (2001a). Random forests. *Machine Learning*, 45, 5–32. doi:10.1023/A:1010933404324

Breiman, L. (2001b). Statistical modeling: The two cultures. *Statistical Science*, 16, 199–231. doi:10.1214/ss/1009213726

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees*. New York, NY: Chapman & Hall.

Britton, B. K., Glynn, S. M., Meyer, B. J., & Penland, M. J. (1982). Effects of text structure on use of cognitive capacity during reading. *Journal of Educational Psychology*, 74, 51–61. doi:10.1037/0022-0663.74.1.51

Burrows, M., & Wheeler, D. J. (1994). *A block sorting lossless data compression algorithm* (Technical Report No. 124). Maynard, MA: Digital Equipment Corporation.

Carnegie Mellon University. (n.d.). *The CMU pronouncing dictionary*. Retrieved from <http://www.speech.cs.cmu.edu/cgi-bin/cmudict>

Carroll, J. B., Davies, P., & Richman, B. (1971). *The American heritage word frequency book*. New York, NY: American Heritage.

Cohen, J., Cohen, P., Aiken, L. S., & West, S. H. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences*. Hillsdale, NJ: Erlbaum.

Cohesion (linguistics). (n.d.). In *Wikipedia*. Retrieved October 1, 2011, from http://en.wikipedia.org/wiki/Cohesion_%28linguistics%29

Collins, M. (2002, July). Discriminative training methods for hidden Markov models: Theory and experiments with perceptron algorithms. In J. Hajic, & Y. Matsumoto (Eds.), *Proceedings of the conference on empirical methods in natural language processing* (pp. 1–8). Philadelphia, PA: Special Interest Group on Linguistic Data and Corpus-Based Approaches to NLP.

Coltheart, M. (1981). The MRC psycholinguistic database. *The Quarterly Journal of Experimental Psychology, Section A: Human Experimental Psychology*, 33, 497–505.

Compton, D. L., Appleton, A. G., & Hosp, M. K. (2004). Exploring the relationship between text-leveling systems and reading accuracy and fluency in second-grade students who are average and poor decoders. *Learning Disabilities Research & Practice*, 19, 176–184. doi:10.1111/j.1540-5826.2004.00102.x

Cook, D., & Swayne, D. F. (2008). *Interactive and dynamic graphics for data analysis with R and Ggobi*. New York, NY: Springer.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41, 391–407. doi:10.1002/(SICI)1097-4571(199009)41:6<391::AID-ASI1>3.0.CO;2-9

Dolch word list. (n.d.). In *Wikipedia*. Retrieved October 1, 2011, from http://en.wikipedia.org/wiki/Dolch_word_list

Duke, N. K. (2000). 3.6 minute per day: The scarcity of informational texts in first grade. *Reading Research Quarterly*, 35, 202–224. doi:10.1598/RRQ.35.2.1

Ehri, L. C., & McCormick, S. (1998). Phases of word learning: Implications for instruction with delayed and disabled readers. *Reading & Writing Quarterly: Overcoming Learning Difficulties*, 14, 135–163. doi:10.1080/1057356980140202

Fitzgerald, J., & Shanahan, T. (2000). Reading and writing relations and their development. *Journal of Educational Psychology*, 92, 3–22. doi:10.1037/0022-0663.92.1.3

Foorman, B. R., Francis, D. J., Davidson, K. G., Harm, M. W., & Griffin, J. (2004). Variability in text features in six Grade 1 basal reading programs. *Scientific Studies of Reading*, 8, 167–197. doi:10.1207/s1532799xssr0802_4

Fountas, I. C., & Pinnell, G. S. (1996). *Guided reading: Good first teaching for all children*. Portsmouth, NH: Heinemann.

Fountas, I. C., & Pinnell, G. S. (2012). Guided reading: The romance and the reality. *The Reading Teacher*, 66, 268–284. doi:10.1002/TRTR.01123

Fry Word List—1,000 High Frequency Words. (2012). In *K12Reader: Reading instruction resources for teachers & parents*. Retrieved from <http://www.k12reader.com/fry-word-list-1000-high-frequency-words/>

Gamson, D. A., Lu, X., & Eckert, S. A. (2013). Challenging the research base of the common core state standards: A historical reanalysis of text complexity. *Educational Researcher*, 42, 381–391. doi:10.3102/0013189X13505684

Gervasi, V., & Ambriola, V. (2003). Quantitative assessment of textual complexity. In L. Merlini Barbaresi (Ed.), *Complexity in language and text* (pp. 199–230). Pisa, Italy: Edizioni Plus.

Graesser, A. C., & McNamara, D. S. (2011). Computational analyses of multilevel discourse comprehension. *Topics in Cognitive Science*, 3, 371–398. doi:10.1111/j.1756-8765.2010.01081.x

Graesser, A. C., McNamara, D. S., & Kulikowich, J. M. (2011). Coh-Metrix: Providing multilevel analyses of text characteristics. *Educational Researcher*, 40, 223–234. doi:10.3102/0013189X11413260

Green, S. B., & Salkind, N. J. (2011). *Using SPSS for Windows and Macintosh: Analyzing and understanding data* (6th ed.). Upper Saddle River, NJ: Prentice Hall.

Grömping, U. (2009). Variable importance assessment in regression: Linear regression versus random forest. *The American Statistician*, 63, 308–319. doi:10.1198/tast.2009.08199

Gusfield, D. (1997). *Algorithms on strings, trees and sequences: Computer science and computational biology*. Cambridge, England: University of Cambridge. doi:10.1017/CBO9780511574931

Halliday, M. A. K., & Hasan, R. (1976). *Cohesion in English*. London, England: Longman.

Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). *The elements of statistical learning* (2nd ed.). New York, NY: Springer. doi:10.1007/978-0-387-84858-7

Heldsinger, S., & Humphry, S. (2010). Using the method of pairwise comparison to obtain reliable teacher assessments. *The Australian Educational Researcher*, 37, 1–19. doi:10.1007/BF03216919

Hiebert, E. H. (2011). Texts for beginning readers: The search for optimal scaffolds. In C. Conrad & R. Serlin (Eds.), *The SAGE handbook for research in education: Pursuing ideas as the keystone of exemplary inquiry* (pp. 413–428). Thousand Oaks, CA: Sage.

Hiebert, E. H. (2012). The common core's staircase of text complexity: Getting the size of the first step right. *Reading Today*, 29, 26–27.

Hiebert, E. H., & Fisher, C. W. (2007). The critical word factor in texts for beginning readers. *The Journal of Educational Research*, 101, 3–11. doi:10.3200/JOER.101.1.3-11

Hiebert, E. H., & Pearson, P. D. (2010). *An examination of current text difficulty indices with early reading texts*. (Reading Research Report No. 10–01). Santa Cruz, CA: TextProject.

Howes, D. H., & Solomon, R. L. (1951). Visual duration thresholds as a function of word probability. *Journal of Experimental Psychology*, 41, 401–410.

Juel, C., & Roper-Schneider, D. (1985). The influence of basal readers on first grade reading. *Reading Research Quarterly*, 20, 134–152. doi:10.2307/747751

Juola, P. (2003). Assessing linguistic complexity. In M. Miestamo, K. Sinne Sinnemäki, & F. Karlsson (Eds.), *Language complexity: Typology, contact, change* (pp. 89–108). Amsterdam, the Netherlands: Benjamins Publishing Co.

Kauffman, S. A. (1995). *At home in the universe: The search for laws of self-organization and complexity*. Oxford, England: Oxford University Press.

Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. Cambridge, England: Cambridge University Press.

Klare, G. R. (1974). Assessing readability. *Reading Research Quarterly*, 10, 62–102. doi:10.2307/747086

Kolen, M. M., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). New York, NY: Springer-Verlag. doi:10.1007/978-1-4757-4310-4

Koslin, B. I., Zeno, S., & Koslin, S. (1987). *The DRP: An effective measure in reading*. New York, NY: College Entrance Examination Board.

Kruskal, J. B. (1999). An overview of sequence comparison. In D. Sankoff & J. B. Kruskal (Eds.), *Time warps, string edits, and macromolecules: The theory and practice of sequence comparison* (pp. 1–44). Stanford, CA: Center for the Study of Language and Information.

Kuperman, V., Stadthagen-Gonzalez, H., & Brysbaert, M. (2012). Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods*, 44, 978–990. doi:10.3758/s13428-012-0210-4

Kusters, W. (2008). Complexity in linguistic theory, language learning and language change. In M. Miestamo, K. Sinnemäki, & F. Karlsson (Eds.), *Language complexity: Typology, contact, change* (pp. 3–22). Amsterdam, the Netherlands: Benjamins. doi:10.1075/slcs.94.03kus

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction and representation of knowledge. *Psychological Review*, 104, 211–240. doi:10.1037/0033-295X.104.2.211

Langer, J. A., Campbell, J. R., Neuman, S. B., Mullis, I. V. S., Persky, H. R., & Donahue, P. S. (1995). *Reading assessment redesigned: Authentic texts and innovative instruments in NAEP's 1992 survey*. Washington, DC: U.S. Department of Education, Office of Educational Research and Improvement.

Levenshtein, V. I. (1965). Binary codes capable of correcting deletions, insertions, and reversals. *Doklady Akademii Nauk SSSR*, 163, 845–848.

Luhn, H. P. (1958). The automatic creation of literature abstracts. *IBM Journal of Research & Development*, 2, 159–165. doi:10.1147/rd.22.0159

Malvern, D. D., Richards, B. J., Chipere, N., & Durán, P. (2009). *Lexical diversity and language development: Quantification and assessment*. New York, NY: Palgrave Macmillan.

Mandler, M. J., & Johnson, N. S. (1977). Remembrance of things parsed: Story structure and recall. *Cognitive Psychology*, 9, 111–151. doi:10.1016/0010-0285(77)90006-8

McNamara, D. S., Graesser, A. C., McCarthy, P. M., & Cai, Z. (2014). *Automated evaluation of text and discourse with Coh-Metrix*. New York, NY: Cambridge University Press.

McNamara, D. S., & Kintsch, W. (1996). Learning from text: Effects of prior knowledge and text coherence. *Discourse Processes*, 22, 247–288. doi:10.1080/01638539609544975

Menon, S., & Hiebert, E. H. (1999). *Literature anthologies: The task for first-graders*. Ann Arbor, MI: Center for the Improvement of Early Reading Achievement.

Merlini Barbaresi, L. M. (2002). Text linguistics and literary translation. In A. Riccardi (Ed.), *Translation studies: Perspectives on an emerging discipline* (pp. 120–132). Cambridge, United Kingdom: Cambridge University Press.

Merlini Barbaresi, L. M. (2003). Towards a theory of text complexity. In L. Merlini Barbaresi (Ed.), *Complexity in language and text* (pp. 23–66). Pisa, Italy: Edizioni Plus.

Mesmer, H. A. (2006). Beginning reading materials: A national survey of primary teachers' reported uses and beliefs. *Journal of Literacy Research*, 38, 389–425. doi:10.1207/s15548430jlr3804_2

Mesmer, H. A., Cunningham, J. W., & Hiebert, E. H. (2012). Toward a theoretical model of text complexity for the early grades: Learning from the past, anticipating the future. *Reading Research Quarterly*, 47, 235–258.

MetaMetrics. (n.d.-a). *Text corpus*. Durham, NC: MetaMetrics.

MetaMetrics. (n.d.-b). *Word corpus*. Durham, NC: MetaMetrics.

Metsala, J. L. (1999). Young children's phonological awareness and non-word repetition as a function of vocabulary development. *Journal of Educational Psychology*, 91, 3–19. doi:10.1037/0022-0663.91.1.3

Miestamo, M. (2009). Implicational hierarchies and grammatical complexities. In G. Sampson, D. Gil, & P. Trudgill (Eds.), *Language complexity as an evolving variable* (pp. 80–97). Oxford, England: Oxford University Press.

Mitchell, T. (1997). *Machine learning*. Columbus, OH: McGraw Hill.

Mohri, M., Rostamizadeh, A., & Talwalkar, A. (2012). *Adaptive computation and machine learning series: Foundations of machine learning*. Cambridge, MA: MIT Press.

MRC Psycholinguistic Database. (n.d.). Retrieved from <http://www.psych.rl.ac.uk>

Muter, V., Hulme, C., Snowling, M. J., & Stevenson, J. (2004). Phonemes, times, vocabulary, and grammatical skills as foundations of early reading development: Evidence from a longitudinal study. *Developmental Psychology*, 40, 665–681. doi:10.1037/0012-1649.40.5.665

Nason, M., Emerson, S., & LeBlanc, M. (2004). CARTscans: A tool for visualizing complex models. *Journal of computational and graphical statistics*, 13, 807–825. doi:10.1198/106186004X11417

National Governors Association Center for Best Practices & Council of Chief State School Officers. (2010). *Common core state standards for English language arts and literacy in history/social studies, science, and technical subjects*. Washington, DC: Author. Retrieved from www.corestandards.org/assets/CCSSI_ELA%20Standards.pdf

Nelson, J., Perfetti, C., Liben, D., & Liben, M. (2011). *Measures of text difficulty: Testing their predictive value for grade levels and student performance* (Technical Report to the Gates Foundation). Retrieved from www.ccsso.org/Documents/2012/Measures%20ofText%20Difficulty_final.2012.pdf

Nerbonne, J., & Heeringa, W. J. (2001). Computational comparison and classification of dialects. *Dialectologia et Geolinguistica*, 9, 69–83.

Paivio, A., Yuille, J. C., & Madigan, S. A. (1968). Concreteness, imagery, and meaningfulness values for 925 nouns [Monograph]. *Journal of Experimental Psychology*, 76(1, Pt. 2), 1–25.

Patton, M. (1990). *Qualitative evaluation research methods*. Beverly Hills, CA: Sage.

Pearson Education. (2014). *Reading Maturity Metric*. Retrieved from <http://www.readingmaturity.com/rmm-web/#/>

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., . . . Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.

Prince, E. (1997). On the functions of the left-dislocation in English discourse. In A. Kamio (Ed.), *Directions in functional linguistics* (pp. 117–144). Amsterdam, the Netherlands: Benjamins. doi:10.1075/slcs.36.08pri

The REAP project: Reader-specific lexical practice for improved reading comprehension. (2011). Retrieved from <http://reap.cs.cmu.edu/>

Renaissance Learning. (2014). *ATOS analyzer*. Retrieved from <http://www.renlearn.com/atos/>

Rescher, N. (1998). *Complexity: A philosophical overview*. London, England: Transaction.

Rosenblatt, L. M. (1938). *Literature as exploration*. New York, NY: Appleton-Century.

Rosenblatt, L. (2005). *Making meaning with texts: Selected essays*. Portsmouth, NH: Heinemann.

Rudrum, D. (2005). From narrative representation to narrative use: Towards the limits of definition. *Narrative*, 13, 195–204. doi:10.1353/nar.2005.0013

Rumelhart, D. E. (1985). Toward an interactive model of reading. In H. Singer & R. B. Ruddell (Eds.), *Theoretical models and processes of reading* (pp. 722–750). Newark, DE: International Reading Association.

Sanders, N. C., & Chinn, S. B. (2009). Phonological distance measures. *Journal of Quantitative Linguistics*, 16, 96–114. doi:10.1080/09296170802514138

Schatschneider, C., Fletcher, J. M., Francis, D. J., Carlson, C. D., & Foorman, B. R. (2004). Kindergarten prediction of reading skills: A longitudinal comparative analysis. *Journal of Educational Psychology*, 96, 265–282. doi:10.1037/0022-0663.96.2.265

Schwanenflugel, P. J., & Akin, C. E. (1994). Developmental trends in lexical decisions for abstract and concrete words. *Reading Research Quarterly*, 29, 250–264. doi:10.2307/747876

Shanahan, T., Fisher, D., & Frey, N. (2012). The challenge of challenging text. *Educational Leadership*, 69, 58–62.

Sheehan, K. M., Kostin, I., Futagi, Y., & Flor, M. (2010, December). *Generating automated text complexity classifications that are aligned with targeted text complexity standards* (ETS RR-10-28). Princeton, NJ: Educational Testing Service.

Shin, J., Deno, S. L., & Espin, C. (2000). Technical adequacy of the maze task for curriculum-based measurement of reading growth. *The Journal of Special Education*, 34, 164–172. doi:10.1177/002246690003400305

Simon, H. A. (1962). The architecture of complexity. *Proceedings of the American Philosophical Society*, 106, 467–482.

Sleator, D., & Temperley, D. (1991, October). *Parsing English with a link grammar* (Carnegie Mellon University Computer Science Technical Report No. CMU-CS-91-196). Pittsburgh, PA: Carnegie Mellon University.

Snow, C. (2002). *Reading for understanding: Toward an R&D program in reading comprehension*. Santa Monica, CA: RAND.

Solso, R. L., Barbuto, P. F., Jr., & Juel, C. L. (1979). Methods & designs: Bigram and trigram frequencies and versatilities in the English language. *Behavior Research Methods & Instrumentation*, 11, 475–484. doi:10.3758/BF03201360

Stanovich, K. E. (1986). Matthew effects in reading: Some consequences of individual differences in the acquisition of literacy. *Reading Research Quarterly*, 21, 360–407. doi:10.1598/RRQ.21.4.1

Steen, G. (1999). Genres of discourse and definition of literature. *Discourse Processes*, 28, 109–120. doi:10.1080/01638539909545075

Stenner, A. J., Burdick, H., Sanford, E., & Burdick, D. (2006). How accurate are Lexile text measures? *Journal of Applied Measurement*, 7, 307–322.

Strobl, C., Malley, J., & Tutz, G. (2009). An introduction to recursive partitioning: Rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological Methods*, 14, 323–348. doi:10.1037/a0016973

Temperley, D., Sleator, D., & Lafferty, J. (2012). *Link grammar*. Retrieved from <http://www.link.cs.cmu.edu/link/>

Vadasy, P. F., Sanders, E. A., & Peyton, J. A. (2005). Relative effectiveness of reading practice or word-level instruction in supplemental tutoring: How text matters. *Journal of Learning Disabilities*, 38, 364–380. doi:10.1177/00222194050380041401

Vanderplas, J., & Connolly, A. (2009). Reducing the dimensionality of data: Locally linear embedding of Sloan galaxy spectra. *The Astronomical Journal*, 138, 1365–1379. doi:10.1088/0004-6256/138/5/1365

van der Sluis, F., & van den Broek, E. L. (2010). Using complexity measures in information retrieval. In *Proceedings of the third symposium on information interaction in context* (pp. 18–22). New Brunswick, NJ: ACM.

Wasserman, L. (2014). Rise of the machines. In L. Xihong, D. L. Banks, C. Genest, G. Molenberghs, D. W. Scott, & J.-L. Want (Eds.), *Past, present and future of statistical science* (pp. 525–536). New York, NY: Taylor & Francis. doi:10.1201/b16720-49

Whaley, J. F. (1981). Readers' expectations for story structures. *Reading Research Quarterly*, 17, 90–114. doi:10.2307/747250

Williamson, G. L. (2008). A text readability continuum for postsecondary readiness. *Journal of Advanced Academics*, 19, 602–632.

Woolams, A. M. (2005). Imageability and ambiguity effects in speeded naming: Convergence and divergence. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31*, 878–890.

Wright, B., & Stone, M. (1999). *Measurement essentials* (2nd ed.). Wilmington, DE: Wide Range.

Yarkoni, T., Balota, D., & Yap, M. (2008). Moving beyond Coltheart's *N*: A new measure of orthographic similarity. *Psychonomic Bulletin & Review, 15*, 971–979. doi:10.3758/PBR.15.5.971

Zhang, Z., & Wang, J. (2007). MLLE: Modified locally linear embedding using multiple weights. In B. Schölkopf, J. Platt, & T. Hoffman (Eds.), *Advances in neural information processing systems 19: Proceedings of the Twentieth Annual Conference on Neural Information Processing Systems* (pp. 1593–1600). Cambridge, MA: MIT Press.

Received December 2, 2013

Revision received May 8, 2014

Accepted June 3, 2014 ■

Successful Learning With Multiple Graphical Representations and Self-Explanation Prompts

Martina A. Rau
University of Wisconsin—Madison

Vincent Aleven
Carnegie Mellon University

Nikol Rummel
Ruhr-Universität Bochum

Research shows that multiple external representations can significantly enhance students' learning. Most of this research has focused on learning with text and 1 additional graphical representation. However, real instructional materials often employ multiple *graphical* representations (MGRs) in addition to text. An important open question is whether the use of MGRs leads to better learning than a single *graphical* representation (SGR) when the MGRs are presented separately, 1-by-1 across consecutive problems, accompanied by text and numbers. A further question is whether providing support for students to relate the different representations to the key concepts that they depict can enhance their benefit from MGRs. We investigated these questions in 2 classroom experiments that involved problem solving practice with an intelligent tutoring system for fractions. Based on 112 sixth-grade students, Experiment 1 investigated whether MGRs lead to better learning outcomes than 1 commonly used SGR, and whether this effect can be enhanced by prompting students to self-explain key concepts depicted by the graphical representations. Based on 152 fourth- and fifth-grade students, Experiment 2 investigated whether the advantage of MGRs depends on the specific representation chosen for the SGR condition because prior research suggests that some SGRs might promote learning more than others. Both experiments demonstrate that MGRs lead to better conceptual learning than an SGR, provided that students are supported in relating graphical representations to key concepts. We extend research on multiple external representations by demonstrating that MGRs (presented in addition to text and 1-by-1 across consecutive problems) can enhance learning.

Keywords: multiple representations, self-explanation prompts, intelligent tutoring systems, fractions, classroom experiment

In the educational psychology literature, there is substantial evidence that multiple external representations can enhance students' learning, compared to a single external representation (Ainsworth, Bibby, & Wood, 2002; Schnottz & Bannert, 2003). By and large, the educational psychology literature on learning with multiple external representations has focused on learning with text and one additional graphical representation (e.g., Ainsworth & Loizou, 2003; Bodemer, Plötzner, Bruchmüller, & Häcker, 2005; Eitel, Scheiter, & Schüler, 2013; Schnottz & Bannert, 2003). The advantage of learning with text and a graphical representation (compared

to learning with text alone) has been attributed to the fact that multiple external representations stimulate deeper processing by requiring learners to integrate information across different representations. However, instructional materials found in real educational settings typically contain multiple *graphical* representations in addition to text (van Someren, Boshuizen, & de Jong, 1998), for instance in math (e.g., Arcavi, 2003), chemistry (e.g., Kozma & Russell, 2005), biology (e.g., Cook, Wiebe, & Carter, 2008), physics (e.g., van der Meij & de Jong, 2006), engineering (e.g., Nathan, Walkington, Srisurikan, & Alibali, 2011), and programming (e.g., Kordaki, 2010). Multiple *graphical* representations are typically used because each individual graphical representation emphasizes a subset of the domain-relevant information and the different graphical representations therefore provide complementary information. For this reason, students are often provided with multiple *graphical* representations in addition to text and numbers, to provide the information necessary to form a coherent mental model of the domain.

We are not aware of experimental studies that investigate whether multiple graphical representations (MGRs) lead to better learning than a single graphical representation (SGR), when each is provided in addition to text and numbers, even though educational psychology research is often conducted in the context of materials that contain MGRs. For example, Nathan et al. (2011)

This article was published Online First July 7, 2014.

Martina A. Rau, Department of Educational Psychology, University of Wisconsin—Madison; Vincent Aleven, Human-Computer Interaction Institute, Carnegie Mellon University; Nikol Rummel, Institute of Education, Ruhr-Universität Bochum.

This work was supported by National Science Foundation Grant REESE-21851-1-1121307 and by the Pittsburgh Science of Learning Center, which is funded by National Science Foundation Grant SBE-0354420.

Correspondence concerning this article should be addressed to Martina A. Rau, Department of Educational Psychology, University of Wisconsin—Madison, 1025 West Johnson, Madison, WI 53706. E-mail: marau@wisc.edu

observed students' ability to coordinate between MGRs across STEM courses via gesture. Kozma and Russell (2005) observed chemistry experts' use of MGRs as they work in chemistry laboratories. Other related research has focused on learning with animations. For example, Scheiter, Gerjets, Huk, Imhof, and Kammerer (2009) contrasted students' learning from a single type of animations (either realistic or schematic) to multiple types of animations (both realistic and schematic). This study did not show an advantage of learning with multiple types of representations, possibly because one type of animation had overwhelmingly positive effects on all target concepts. In other words, the different types of animations failed to produce complementary benefits on students' learning, so that multiple types of animations were not more effective than a single type of animation. Taken together, these prior studies do not systematically investigate whether instruction that uses MGRs is *more effective* than instruction that uses only an SGR. In other words, in spite of the widespread use of MGRs, it remains an open question whether prior research on learning with text and graphic generalizes to learning with multiple *graphical* representations, when each is presented in addition to text and numbers.

There are many possible ways of presenting MGRs to students: We can present them concurrently (i.e., presenting more than one graphical representation at the same time) or consecutively (i.e., presenting one graphical representation at a time, but different graphical representations across a sequence of problems). In this article, we focus on the question whether MGRs lead to better learning than an SGR *when presented consecutively* across a sequence of problems. Presenting MGRs consecutively is a logical next step when extending research on learning with multiple external representations to learning with multiple graphical representations: It allows us to investigate whether increasing the number of graphical representations leads to better learning without increasing the number of connections students have to make between representations (because in each problem, students can only connect one graphical representation to text and numbers but cannot connect different graphical representations to one another). By contrast, concurrent presentation of MGRs may place very high demands on students' cognitive load due to many possible connections between the different graphical representations, text, and numbers. Furthermore, consecutive presentation of MGRs is common practice in many educational materials. The central hypothesis of this article is that MGRs, when presented consecutively, will enhance students' learning because students can form a more accurate mental model of the domain knowledge by gradually refining it across a sequence of problems. This refinement process is facilitated by MGRs because they emphasize complementary conceptual aspects of the domain knowledge.

Theoretical Perspectives on Learning With Multiple External Representations

Current theoretical frameworks for learning with multiple external representations do not address the question of whether MGRs lead to better learning than an SGR. According to the *cognitive theory of multimedia learning* (CTML; Mayer, 2003, 2005), the advantage of learning with text and graphic can be attributed to the more efficient use of working memory capacity and to deeper conceptual integration of the learning material. The

CTML assumes that verbal and pictorial information are processed in different information channels. Even though text is often presented visually (i.e., in written form), it is encoded into a verbal model within working memory, whereas pictorial information is encoded visually. Because the capacity of each part of working memory (a visual channel and a verbal channel) is limited but additive (Sweller, van Merriënboer, & Paas, 1998), learning with both text and graphic makes better use of the learner's working memory capacity than learning with text alone. Furthermore, active integration of the verbal model and the pictorial model into one coherent mental model requires deep conceptual processing of the content, which enhances learning, compared to text alone.

The CTML does not specifically address the use of MGRs presented consecutively, accompanied by text. Learning materials with an SGR or with MGRs (when each is presented individually in addition to text and numbers) both use the same number of information channels. The CTML might predict that MGRs lead to cognitive overload in the pictorial part of working memory if they were presented concurrently, which is known to hamper learning (Sweller et al., 1998). On the other hand, based on the CTML, one might predict that MGRs will be more effective than an SGR when they are presented consecutively by enhancing active integration and deeper conceptual processing of the learning content, because students may gradually refine their mental model based on the different conceptual aspects that each graphical representation emphasize. Specifically, each time students encounter a different graphical representation, they may incorporate the conceptual perspective emphasized in this graphical representation into their mental model of the learning content.

A second relevant theory, the *integrated model of text and picture comprehension* (ITPC; Schnottz, 2005; Schnottz & Bannert, 2003), also does not explicitly address the use of MGRs. Under this theory, the advantage of learning with multiple external representations stems from the fact that text and graphic lead to different internal representations. During mental model formation, students integrate these internal representations via structure mapping (Gentner, 1983). The resulting deep integration across representations accounts for the effectiveness of text and graphic over text alone. Based on the ITPC, one might predict that MGRs enhance learning because they provide complementary information, which allows students to form more accurate mental models. However, the ITPC also states that understanding each representation creates cognitive costs (Schnottz, 2005). Another possible interpretation of the ITPC is therefore that MGRs do not lead to better learning than an SGR as the costs of understanding each graphical representation may not outweigh the benefit of integrating complementary information into a mental model.

Taken together, neither the CTML nor the ITPC make specific predictions as to whether MGRs lead to better learning than an SGR, even though both theoretical frameworks are consistent with the idea that MGRs might enhance learning by allowing students to form more sophisticated mental models of domain-relevant concepts. Therefore, the goal of the present article is to close the gap between prior educational psychology research that has mainly focused on learning with text and graphic and the common practice to include MGRs, in addition to text-based representations.

Multiple Graphical Representations of Fractions

We address this question in the important but challenging domain of fractions learning, one of the many domains in which MGRs (e.g., circles, rectangles, and number lines) are used extensively (National Mathematics Advisory Panel [NMAP], 2008). In spite of the extensive use of MGRs in fractions instruction, the advantage of MGRs of fractions over an SGR has not yet been systematically investigated. Commonly used U.S. middle-school mathematics curricula (e.g., Bennett, 2004; Fitzgerald, Lappan, & Fey, 2004; Hake, 2004) employ a wide variety of graphical representations of fractions, such as area models (e.g., circles, rectangles), linear models (e.g., number lines, or liquid container models), and discrete models (e.g., sets of objects). These graphical representations emphasize different conceptual interpretations of fractions, such as fractions as parts of a whole in area models, fractions as measurements in linear models that show fractions as a segment of a length that can be measured, and fractions as ratios in discrete models that depict fractions as a subset of objects with a certain property out of a larger set of objects (Charalambous & Pitta-Pantazi, 2007). The use of multiple graphical representations of fractions in instructional materials has been shown to be effective in observational studies (e.g., Moss & Case, 1999) and in case studies (e.g., Kafai, Franke, Ching, & Shih, 1998). However, these studies did not systematically compare learning with an SGR to learning with MGRs. Thus, in addition to being of theoretical relevance, the choice of fractions as a domain for our study enhances the practical relevance of this research.

We investigated these questions within the context of the Fractions Tutor (Rau, Aleven, Rummel, & Rohrbach, 2013). The Fractions Tutor is a type of Cognitive Tutor (Koedinger & Corbett, 2006), namely, an example-tracing tutor (Aleven, McLaren, Sewall, & Koedinger, 2009). Like all Cognitive Tutors, it supports tutored problem solving, providing step-by-step guidance as students solve complex problems. The use of a Cognitive Tutor as research platform is attractive because Cognitive Tutors support learners in a common instructional scenario, namely, problem-solving practice, they have a proven track record in improving students' mathematics achievement (Koedinger & Aleven, 2007; Pane, Griffin, McCaffrey, & Karam, 2013), and they are being used in a large number of classrooms across the United States, about 600,000 students yearly (Koedinger & Corbett, 2006). Furthermore, Cognitive Tutors can support *interactive* graphical representations of fractions and provide targeted feedback on students' use of these interactive representations as they solve fractions problems.

Scaffolding Learning With Multiple Representations Through Self-Explanation Prompts

Providing learners with multiple representations does not automatically result in better learning (Ainsworth et al., 2002). As discussed, in order to benefit from multiple representations, learners must conceptually understand how the different representations depict key concepts and integrate these concepts into a coherent mental model of the domain (Ainsworth, 2006; de Jong, et al., 1998). However, students tend not to spontaneously engage in such sense-making activities (Ainsworth, 2006; Yerushamly, 1991), which can hamper their learning of domain knowledge

(Ainsworth et al., 2002; de Jong, et al., 1998; Gobert et al., 2011; Gutwill, Frederiksen, & White, 1999; van der Meij & de Jong, 2006).

Several studies lead to the hypothesis that integration processes involved in learning with multiple representations can happen through self-explanation activities. Self-explanation activities are explanations to oneself that elaborate information provided in the learning materials, make connections to prior knowledge, and refine mental models (Chi, Bassok, Lewis, Reimann, & Glaser, 1989; Wylie & Chi, in press). Research shows that students who generate a larger number of high-quality self-explanations show the largest benefits from multiple external representations (Ainsworth & Loizou, 2003). However, students tend not to spontaneously engage in high-quality self-explanation activities (Ainsworth & Loizou, 2003). This observation leads to the hypothesis that prompting students to self-explain how different representations relate to the conceptual aspects of the domain may further enhance their benefit from multiple representations. Berthold, Eysink, and Renkl (2009) found that prompting students to self-explain while studying multirepresentational learning materials had a positive effect on both conceptual and procedural knowledge. Zhang and Linn (2011) found positive effects of enhancing student-generated explanations while learning with dynamic chemistry representations. In a recent review of the effects of self-explanation prompts on learning in multimedia environments, Wylie and Chi (in press) argued that environments in which students have to relate multiple representations, self-explanation prompts can be an effective instructional strategy. However, we are not aware of a systematic investigation of the effect of self-explanation prompts on students' benefit from multiple representations *compared to* a single representation. In line with prior research, we expect that prompting students to self-explain will increase the likelihood that they will engage in deeper sense-making activities with MGRs, thus increasing their benefit from MGRs.

In our research, we use a focused form of self-explanation support; namely, menu-based prompts. In complex multimedia environments, such focused forms of supporting self-explanation have been hypothesized to be more effective than traditional open-ended approaches (Wylie & Chi, in press). Supporting self-explanation by the means of menu-based selections is a type of support chosen in many empirical studies with Cognitive Tutors (see Aleven & Koedinger, 2002; Atkinson, Renkl, & Merrill, 2003) and was more effective than open-ended self-explanation prompts in several studies (Gadgil, Nokes-Malach, & Chi, 2012; Johnson & Mayer, 2010; van der Meij & de Jong, 2011).

Overview of Experiments

We conducted two classroom experiments to investigate whether MGRs presented consecutively lead to better learning than one SGR. In both conditions, each graphical representation was accompanied by text and numbers. Further, Experiment 1 investigates whether self-explanation prompts enhance this hypothesized advantage of MGRs.

Experiment 1

Classroom Experiment 1 investigated, in a 2×2 design, the effects of learning with MGRs compared to learning with an SGR

and the effects of being prompted to self-explain compared to not being prompted. In all conditions, students worked with only one graphical representation per tutor problem (presented in addition to text and numbers). That is, students in the SGR conditions encountered only one graphical representation across all tutor problems, namely, the number line. By contrast, students in the MGR conditions encountered different graphical representations across consecutive tutor problems. We chose the number line representation for the SGR conditions because number lines are considered a privileged representation that relates to many math concepts, such as integers and decimals, and that provides a foundation for algebra (Siegler et al., 2010).

Specifically, we investigated the following hypotheses:

Hypothesis 1.1: Students who learn with MGRs will outperform students who learn with an SGR on all measures of robust learning; namely, reproduction and transfer of conceptual knowledge, and reproduction and transfer of procedural knowledge (main effect of number of graphical representations).

Hypothesis 1.2: Students who receive self-explanation prompts will outperform students who do not receive such prompts on all measures of robust learning (main effect of self-explanation prompts).

Hypothesis 1.3: Students who learn with MGRs will outperform students who learn with an SGR *in particular* when they receive self-explanation prompts on all measures of robust learning (interaction between number of graphical representations and self-explanation prompts).

Method

Participants. One hundred thirty-two sixth-grade students from a U.S. middle school participated in the study during regular mathematics instruction. The school district was among the 15% lowest ranked of 500 Pennsylvania public school districts in the school year of 2007/2008. In the school year of 2007/2008, about half of all students in the school district were enrolled in free or reduced-price lunch programs, roughly two thirds of all students were White, and around one third were African American.¹ Students were aged 10 to 13 years.

Fractions tutors. The tutor used in the study included five different graphical representations of fractions, shown in Figure 1. All students worked through two topics of the Fractions Tutor: equivalent fractions and fraction addition (see Appendix for a description of the topics covered). Figure 2 shows an example of a fraction addition problem with and without self-explanation prompts. As is typical of Cognitive Tutors (e.g., Koedinger & Corbett, 2006), the Fractions Tutor provides problem-solving activities while giving error feedback on all steps. Error feedback was designed to encourage students to reconsider their answer by reminding them of a previously introduced principle, or by providing them with an explanation of their error. At any time, students could request hints that provided guidance on how to solve the next step.

Test instruments. We assessed students' knowledge of fractions three times: prior to the tutoring sessions with a short prior knowledge test and twice after the tutoring sessions with equiva-

lent immediate and delayed posttests. The delayed posttest was administered 1 week after the immediate posttest. Two equivalent posttest forms were created such that the test items were structurally the same, but with different numbers. The order of test forms was counterbalanced. Test items were adapted from standardized state tests and from the fractions literature. We used four measurement scales to assess students' robust knowledge, validated by a confirmatory factor analysis. The four scales differed in whether they tested *reproduction* or *transfer* of fractions knowledge, and whether they tested *conceptual* knowledge or *procedural* knowledge. The conceptual reproduction scale of the test included equivalent fractions problems with the same representations used in the tutor, and items that required students to draw the graphical representations they had seen in the tutor. Procedural reproduction items included equivalent fractions and fraction addition problems that were purely symbolic. Conceptual transfer items included equivalent fractions problems and identifying fractions problems using unfamiliar graphical representations and cover stories not covered in the tutor. The procedural transfer scale included fraction addition problems with unfamiliar graphical representations and fraction subtraction problems (subtraction was not covered by tutor). Example test items are provided in the Appendix. As the purpose of the prior knowledge test was to control for differences in students' prior knowledge rather than to assess students' learning gains, it was a shorter version of the posttests and included only reproduction items. The prior knowledge test had 13 items, the posttests had 18 items. For questions that required multiple steps, partial credit was given for each correct step. The scores reported here are relative scores (i.e., ranging from 0 to 1).

Experimental design. Students were randomly assigned to one of four conditions, which varied on two experimental factors: number of graphical representations (SGR vs. MGRs) and self-explanation prompts (SE vs. noSE). Students in the SE conditions were prompted by the tutor to self-explain what aspects of the given graphical representations correspond to the concepts of numerator and the denominator of the fraction (e.g., "How does the number line show the numerator of the fraction?"), or how the procedure they performed symbolically corresponds to the manipulation of the graphical representations (e.g., "How did you convert the fraction in the circle?"). Students selected their answer from a drop-down menu (see Figure 2). Students in the noSE conditions received the same tutor problems without the prompts.

In the SGR conditions, all problems involved an interactive number line representation (see Figure 2). In the MGR conditions, students worked with the five graphical representations shown in Figure 1, which were presented in an interleaved fashion, so that only one graphical representation was presented at a time, but consecutive problems used different graphical representations. Students first solved a fractions problem using the number line. They then performed the same steps symbolically. Next, students revisited the same problem they had solved with the number line with the four remaining graphical representations. Students in the SE conditions were asked to reflect on what aspect in the graphical representation corresponds to numerator and denominator with

¹ The precise numbers are withheld to preserve anonymity of the participating school.

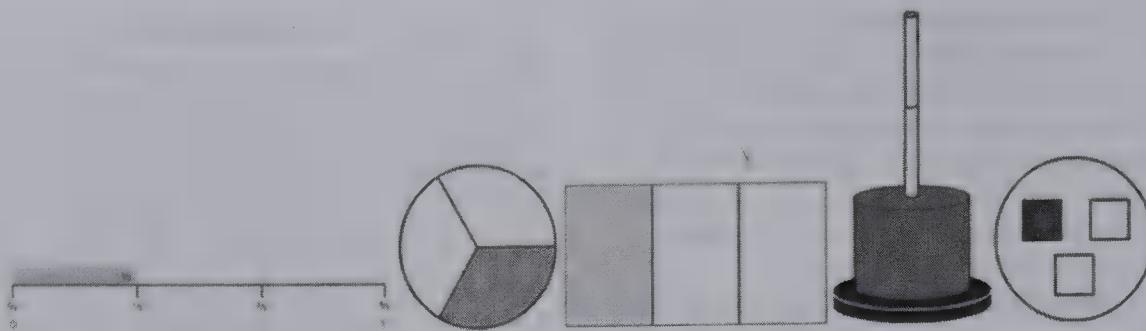


Figure 1. Number line, pie chart, rectangle, stack, set (from left to right) as used in the study.

regard to the steps they previously completed with the number line representation. Students in the noSE conditions skipped this step.

Experimental procedure. On the first day of the study, students completed the prior knowledge test, which took about 20 min. On the next day, students started working with the tutor. They worked with the tutor during their regular mathematics instruction in their school's computer lab for a total of 2.5 hr, spread across two consecutive days. Students worked at their own pace, but time spent with the tutor was held constant across experimental conditions, so that students completed as many tutor problems as they could in the available time. Immediately after finishing the work on the tutor, students completed the immediate posttest, which took about 30 min. Six days later, students completed the delayed posttest.

Results

We excluded students from the analysis if they had been absent for at least two study days, if they were statistical outliers on test performance (i.e., if they performed more than two standard deviations better or worse than their classmates on both the immediate and the delayed posttest), or if they were statistical outliers with respect to the time they spent on the Fractions Tutor (i.e., if the time they spent on the tutor was over two standard deviations more or less than average due to absenteeism or unsupervised work with the tutor outside of class). Data from $N = 112$ students

were included in the data analysis ($n = 29$ in the SGR-noSE condition, $n = 29$ in the SGR-SE condition, $n = 28$ in the MGR-noSE condition, and $n = 26$ in the MGR-SE condition). The number of excluded students did not differ between conditions, $\chi^2(3, N = 21) < 1$. There were no significant differences between conditions on the prior knowledge test ($F < 1$). However, since scores on the prior knowledge test correlated significantly with overall performance in the immediate and delayed posttest ($ps < .01$), we include the prior knowledge test as a covariate in subsequent analyses.

We follow Cohen (1988) and consider an effect size partial η^2 of .01 to be a small effect, .06 a medium effect, and .14 a large effect. Similarly, we consider an effect size d of .20 to be a small effect, .50 a medium effect, and .80 a large effect. All p -values for post hoc comparisons were adjusted using the Bonferroni correction.

We conducted repeated-measures analyses of covariance (ANCOVAs) with students' scores on the prior knowledge test as a covariate, immediate and delayed posttest scores as dependent variables and number of representations and self-explanation prompts as independent variables. In addition, we computed a priori contrasts on the effect of number of representations within the SE conditions and within the noSE conditions to clarify the predicted interaction effect. To clarify the results from the ANCOVAs, we used post hoc comparisons. Adjusted means and

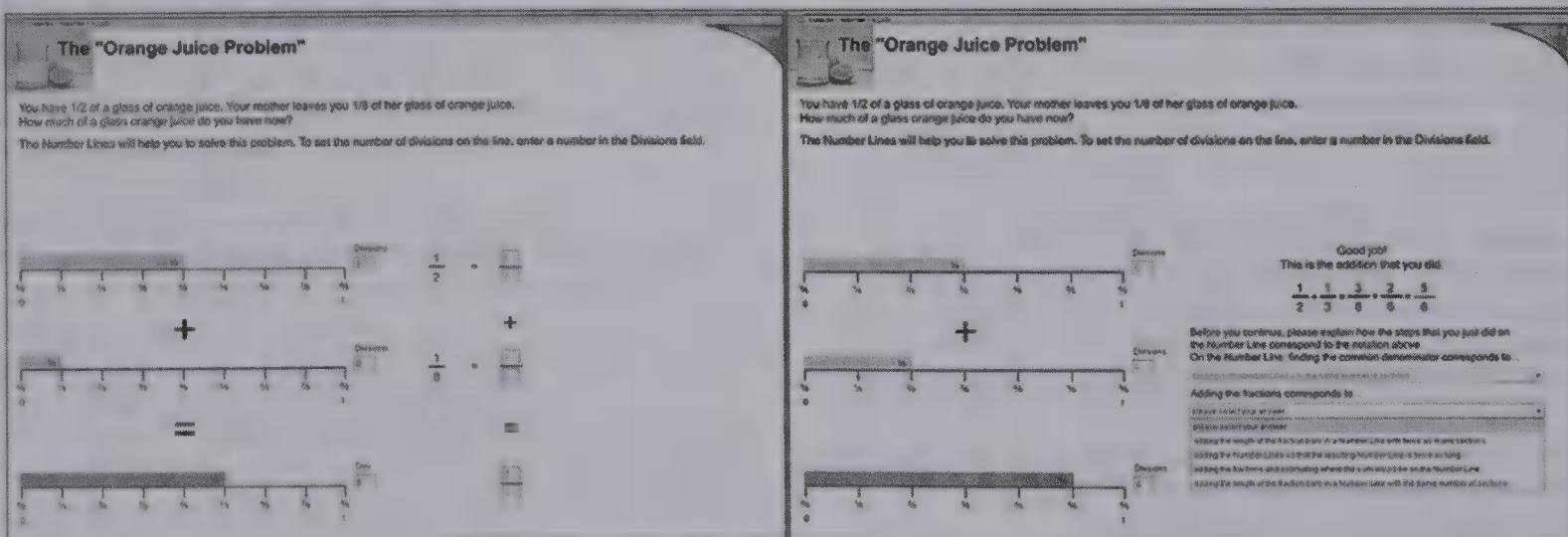


Figure 2. Fraction addition with the number line representation, without self-explanation prompts (left) versus with self-explanation prompts (right).

standard deviations can be found in the Appendix. Table 1 gives an overview of the main effects and interaction effects. Table 2 shows the results from *a priori* contrasts and post hoc comparisons.

To investigate Hypothesis 1.1 (that students who learn with MGRs will outperform students who learn with an SGR), we computed the main effect for number of graphical representations. We found no main effect for the number of graphical representations on any knowledge type ($F_s < 1$). To examine Hypothesis 1.2 (that students benefit from self-explanation prompts), we computed the main effect of self-explanation prompts. We found a significant main effect in favor of the SE conditions on conceptual reproduction, $F(1, 108) = 9.13, p < .01$, partial $\eta^2 = .08$, but not with respect to other knowledge types. Finally, we investigated Hypothesis 1.3 (that self-explanation prompts enhance students' benefit from MGRs) by computing the interaction effect between number of graphical representations and self-explanation prompts. As expected, we found significant interaction effects between the number of graphical representations and self-explanation prompts on conceptual reproduction, $F(1, 108) = 13.02, p < .01$, and procedural transfer, $F(1, 108) = 11.35, p < .01$. To better understand the interaction effect, we computed *a priori* contrasts. Within the SE conditions, we found a significant advantage of the MGR-SE condition over the SGR-SE condition on procedural transfer at the immediate posttest, $t(108) = 2.01, p < .05, d = 0.73$, and marginally significant effects on conceptual reproduction at the immediate posttest, $t(108) = 1.58, p < .10, d = 0.44$, and on conceptual transfer at the delayed posttest, $t(108) = 1.64, p < .10, d = 0.46$. Within the noSE conditions, there was a significant advantage of the SGR-noSE condition over the MGR-noSE condition on procedural transfer at the immediate posttest, $t(108) = 2.80, p < .01, d = 0.99$, but not on any other knowledge type.

To find out why there was no overall advantage of self-explanation prompts for knowledge types other than conceptual reproduction, we used post hoc comparisons. Within the MGR conditions, we found a significant advantage for self-explanation prompts on conceptual reproduction at the immediate and delayed posttests ($ps < .01$), conceptual transfer at the delayed posttest ($p < .05$), and procedural transfer at the immediate posttest ($p < .01$), and a marginal advantage of self-explanation prompts on procedural reproduction at the delayed posttest ($p < .10$). Within the SGR conditions, there were no significant effects of self-explanation prompts.

Finally, we used post hoc comparisons to contrast the two most successful conditions: SGR-noSE and MGR-SE. We found marginal differences on conceptual reproduction and conceptual transfer at the delayed posttest ($ps < .10$) in favor of the MGR-SE condition.

Discussion

Our results do not support Hypothesis 1.1, that students who learn with MGRs will outperform students who learn with an SGR regardless of whether self-explanation prompts are provided. They provide partial support for Hypothesis 1.2, that students who receive self-explanation prompts outperform students who do not receive such prompts. The results support Hypothesis 1.3, that there is an interaction between the number of graphical representations and self-explanation prompts, such that students benefit from MGRs in particular when they are prompted to self-explain the relation between the graphical representations and key concepts of fractions (see Tables 1 and 2).

Let us first consider the effect of number of representations. Our results do not support Hypothesis 1.1: There was no main effect of number of representations (SGR vs. MGRs). Our experiment does not confirm that MGRs *overall* lead to better learning than SGRs. However, in line with Hypothesis 1.3, the results suggest that MGRs enhance learning compared to an SGR when they are accompanied by self-explanation prompts. First, the MGR-SE condition outperformed the SGR-SE condition on conceptual reproduction and conceptual transfer, although the difference was only marginally statistically significant. Second, the MGR-SE condition outperformed the SGR-noSE condition on conceptual reproduction and conceptual transfer. Thus, the effect of MGRs combined with self-explanation prompts over the SGR conditions was specifically found on conceptual knowledge. The evidence is somewhat tentative: Although the interaction effect was statistically significant, some of the post hoc comparisons interpreting this effect were only marginally statistically significant.

Why might MGRs not lead to an *overall* difference on conceptual knowledge, compared to an SGR (Hypothesis 1.1)? In line with the literature on learning with multiple external representations (Ainsworth & Loizou, 2003), it may be that learning with MGRs, especially when they emphasize somewhat disparate conceptual viewpoints, will not be effective unless crucial sense-making processes are supported. Self-explanation prompts appear

Table 1
Overview of Main Effects and Interaction Effects From Experiment 1

Effect	Conceptual knowledge						Procedural knowledge			
	Reproduction		Transfer		Reproduction		Transfer			
	Direction	Significance	Direction	Significance	Direction	Significance	Direction	Significance	Direction	Significance
Main effect of multiple vs. single representations		ns		ns		ns		ns		ns
Main effect of self-explanation prompts	SE > noSE	$p < .01$			ns		ns			ns
Interaction between these two factors	(see Table 2)	$p < .01$			ns		ns		(see Table 2)	$p < .01$

Note. SE = self-explanation.

Table 2
Overview of *A Priori* Contrasts and Post Hoc Comparisons From Experiment 1

Effect	Posttest time	Procedural knowledge									
		Conceptual knowledge					Transfer				
		Reproduction		Transfer		Reproduction		Transfer		Reproduction	
Direction	Significance	Direction	Significance	Direction	Significance	Direction	Significance	Direction	Significance	Direction	Significance
Effect of number of representations with self-explanation prompts	Immediate	MGR-SE > SGR-SE <i>p</i> < .10		MGR-SE > SGR-SE <i>p</i> < .10						MGR-SE > SGR-SE <i>p</i> < .05	
	Delayed	MGR-SE > SGR-SE <i>p</i> < .10									<i>ns</i>
Effect of number of representations without self-explanation prompts	Immediate										
	Delayed										
Effect of self-explanation prompts when working with an MGR	Immediate	MGR-SE > MGR-noSE <i>p</i> < .01		MGR-SE > MGR-noSE <i>p</i> < .01							
	Delayed	MGR-SE > MGR-noSE <i>p</i> < .01									
Effect of self-explanation prompts when working with an SGR	Immediate										
	Delayed										
MGR-SE vs. SGR-noSE	Immediate	MGR-SE > SGR-noSE <i>p</i> < .10		MGR-SE > SGR-noSE <i>p</i> < .10							
	Delayed										

Note. MGR = multiple graphical representation; SE = self-explanation; SGR = single graphical representation.

to be one effective means to do so. This interpretation is in line with the CTML (Mayer, 2005) and the ITPC (Schnitz, 2005), which state that additional representations may create cognitive costs, and that the advantage of multiple representations depends on students' ability to integrate them into a coherent mental model. Our results indicate that self-explanation prompts are a successful means to help students overcome potential costs of MGRs.

Contrary to our hypotheses, we do not find a lasting advantage of MGRs over SGRs on most measures of procedural knowledge. We found significant differences in favor of MGRs on procedural transfer, but this advantage was of temporary nature (i.e., it occurred on the immediate posttest but not on the delayed posttest). The finding that MGRs promote conceptual learning but not (or to a lesser extent) procedural learning is consistent with a view of MGRs as providing complementary conceptual viewpoints. Apparently, MGRs play a lesser role in supporting students' ability to apply a known procedure to solve a familiar task type. MGRs may not help students to perform a procedure *per se*, but rather, in acquiring flexibility to apply a procedure to multiple situations, as supported by the advantage of MGRs (with self-explanation prompts) on procedural transfer.

Taken together, the results from Experiment 1 extend previous research on learning with multiple representations. Most prior research has focused on learning with multiple *external* representations (e.g., text and graphic). Extending this prior research, we demonstrate benefits of using MGRs compared to an SGR, when MGRs are presented one-by-one across problems. Although the evidence is not uniformly strong, overall, a fair summary of the evidence is that students benefit from MGRs, compared to an SGR, provided they are prompted to self-explain how the representations relate to key concepts in the domain. The benefit of learning with MGRs over learning with an SGR was particularly pronounced for conceptual knowledge and persisted until at least 1 week after the intervention.

Several open questions arise from Experiment 1. First, one may ask whether the advantage MGRs over an SGR were due to the choice of graphical representation for the SGR group, namely, the number line. Although the number line considered a powerful representation for fractions (Siegler et al., 2010), it is also the representation students struggle with most (NMAP, 2008). Area models (i.e., circles and rectangles) are considered to be more intuitively accessible (Cramer, Wyberg, & Leavitt, 2008). It is therefore possible that students will benefit equally from a version of the Fractions Tutor that contains only a circle or only a rectangle representation as from the MGR version. Second, one might ask whether having students in the MGR conditions revisit *the same* numerical problems is pedagogically realistic. Might MGRs lead to even better learning if they were presented across *different* numerical problems?

Experiment 2

We conducted a second classroom experiment to investigate whether the advantage of learning with MGRs over learning with an SGR (when provided with self-explanation prompts) is due to the specific SGR used in Experiment 1, as well as to test whether the advantage of MGRs over an SGR could be replicated with an updated version of the Fractions Tutor that includes a more comprehensive curriculum and in which all numerical problems are

different (i.e., no repeats as in Experiment 1). We included three SGR-SE conditions that use only a number line, only a circle, or only a rectangle. We included self-explanation prompts in all conditions, given the conclusion from Experiment 1 that self-explanation prompts enhance students' learning from MGRs.

Specifically, we investigate the following hypothesis: Working with the MGR-SE version of the tutor leads to higher learning gains than working with the SGR-SE version on all measures of robust learning.

Method

Participants. Two hundred and fifty-nine fourth- and fifth-grade students from six different schools in three school districts (31 classes) participated in the study. The schools' rankings in the school year of 2009/2010 were in the top 10% of 2,468 Pennsylvania public schools.² In the school year of 2009/2010, 10%–30% of the students in the participating school districts were enrolled in free or reduced-price lunch programs, over 90% were White, less than 5% African American.

Fractions tutor. We revised the Fractions Tutor in line with our goal to have a more comprehensive tutor curriculum (Rau et al., 2013; see the Appendix). An important change in the Fractions Tutor regards the choice of graphical representations. We decided to include the number line representation, the circle, and the rectangle representation (see Figure 3) but to exclude the set representation.

We excluded the set representation from the Fractions Tutor because the new version covered several topics (e.g., improper fractions, fraction addition) in which the use of sets is not advisable from an educational standpoint, and our experimental design required combining each representation with all topics.

As in Experiment 1, the Fractions Tutor provided problem-solving support in the form of error feedback and hints. In a spiraling curriculum, it covered a sequence of topics three times (see Appendix). As mentioned, self-explanation prompts were included in each tutor problem to help students reflect on the conceptual aspects demonstrated by the graphical representation.

Test instruments. We assessed students' knowledge of fractions three times: immediately before and after using the tutor, and a week later. We created three equivalent test forms (i.e., forms with structurally identical test items that use different numbers) and counterbalanced the order in which they were administered. We made changes to the test used in Experiment 1 in accordance to the changes made to the Fractions Tutor (i.e., addition of tutor topics and choice of graphical representations). The test consisted of 18 items, each of which was worth one point. For questions that required multiple steps, partial credit was given for each correct step. The scores reported here are relative scores (i.e., ranging from 0 to 1). The theoretical structure of the test resulted from a factor analysis performed on the pretest data. Four knowledge types were identified through this factor analysis: reproduction with area models (i.e., problems that involved circles and rectangles), reproduction with the number line, conceptual transfer and procedural transfer. Both reproduction scales included identifying fractions given a graphical representation, making a graphical representation given a symbolic fraction, and recreating the unit given a graphical representation of fractions. Conceptual transfer items included proportional reasoning questions with and without

graphical representations. Procedural transfer items included comparison questions with and without graphical representations. Example test items for each scale are provided in the Appendix.

Experimental design. Students were randomly assigned to either the MGR-SE condition or the SGR-SE condition. Within the SGR-SE condition, we randomly assigned students to either a number-line-only, rectangle-only, or circle-only version of the Fractions Tutor. Students in the MGR-SE condition worked with all three graphical representations in a one-per-problem fashion. Within the MGR-SE condition, we randomly assigned students to one of six possible orders of graphical representations (i.e., number line–rectangle–circle, number line–circle–rectangle, rectangle–number line–circle, rectangle–circle–number line, circle–number line–rectangle, or circle–rectangle–number line) to counterbalance potential order effects. As mentioned, the different graphical representations in the MGR-SE group were presented in an interleaved fashion. This experiment included a number of additional conditions, as reported elsewhere (Rau, Rummel, Aleven, Pacilio, & Tunc-Pekkan, 2012), in which we presented MGRs in different sequences. For the purpose of investigating the advantage of MGRs over SGRs, we chose the condition with an interleaved sequence because it corresponds closest to the procedure in Experiment 1.

Experimental procedure. On the first study day, students completed a pretest, which took about 30 min. On the next day, students started working with the Fractions Tutor. As in Experiment 1, students worked on the Fractions Tutor in the computer lab at their schools for a total of 5 hr during their regular mathematics instruction for five to six consecutive school days (depending on the school's class periods). Students worked with the tutor at their own pace, but the time students spent with the tutor was held constant across classrooms and across experimental conditions, such that the number of problems each student completed was allowed to differ between students. On the day following the tutoring sessions, students took the immediate posttest, which took about 30 min to complete. Seven days after the posttest, students completed an equivalent delayed posttest.

Results

We excluded students who did not encounter all topics covered by the Fractions Tutor. As the Fractions Tutor looped through the sequence of topics three times, we had to exclude students who completed less than 33% of all tutor problems to ensure that all students had encountered each topic of the Fractions Tutor at least once. We further excluded students who missed at least one test day. Due to relatively high rates of absenteeism during regular class time, this results in a total of $N = 152$ with $n = 71$ students in the MGR-SE condition and $n = 81$ students in the SGR-SE condition (thereof $n = 26$ in the number-line-only condition, $n = 25$ in the rectangle-only condition, and $n = 30$ in the circle-only condition). There was no significant difference between the SGR conditions with respect to the number of students excluded ($\chi^2 < 1$), or between the SGR and MGR conditions, $\chi^2(1, N = 259) = 1.579, p > .10$. There were no significant differences between students who were included or excluded on any dependent mea-

² The precise numbers are withheld to preserve anonymity of the participating schools.

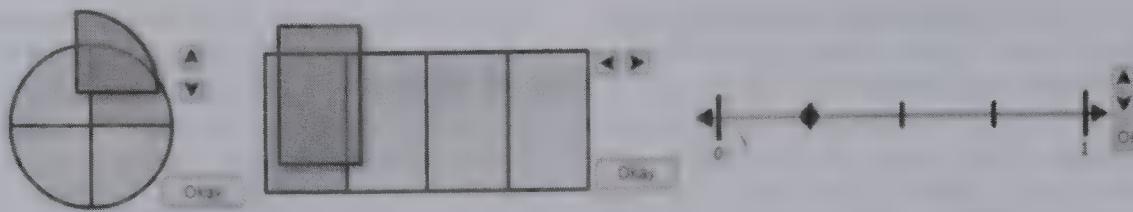


Figure 3. Interactive representations used in fractions tutor: circle, rectangle, and number line.

sure at the pretest ($ps > .10$). There were also no significant differences between conditions at pretest for any dependent measure ($ps > .10$).

We used two different sets of analysis of variance (ANOVA) models to test our hypothesis. To analyze differences in students' learning gains from working with the Fractions Tutor, we computed ANOVAs with time of measurement (pretest, immediate posttest, and delayed posttest) as within-subject factor, number of representations (SGR vs. MGRs) as between-subjects factor, and number of representations by test time as an interaction factor. Within this ANOVA model, we conducted pairwise comparisons for the learning gains from the pretest to the immediate posttest and from the pretest to the delayed posttest, separately for each condition. Table 3 provides an overview of the main effects and a priori comparisons for the ANOVA model. To analyze the differences between conditions, we computed ANCOVAs with number of representations as between-subjects factor, posttest time (immediate and delayed posttest) as within-subject factor, pretest scores as covariates and the immediate and the delayed posttest as repeated, dependent measures. Within these ANCOVA models, we used a priori contrasts to compare the MGR-SE and SGR-SE conditions at the immediate posttest and at the delayed posttest separately. For each model, dependent measures were students' scores on the pretest, the immediate posttest, and the delayed posttest on reproduction with number lines, reproduction with area models, conceptual transfer, and procedural transfer, respectively. All reported p -values were adjusted using the Bonferroni correction. Table 4 shows the results from the ANCOVA model. We provide the estimated means and standard deviations by condition and test time in the Appendix.

Learning effects. We first explored overall learning gains using the ANOVA model, computing the main effect of test time on students' scores at the pretest, the immediate posttest, and the delayed posttest. The effect of test time was significant on reproduction with area models, $F(2, 445) = 3.59, p < .05$, partial $\eta^2 = .01$, reproduction with the number line, $F(2, 445) = 9.25, p < .01$, partial $\eta^2 = .04$, and conceptual transfer, $F(2, 445) = 4.55, p < .01$, partial $\eta^2 = .02$, such that students' scores were higher on the posttests than on the pretest.

Differences between conditions. As a first step, we tested whether the different SGR-SE conditions (i.e., the number-line-only condition, the rectangle-only condition, and the circle-only condition) could be treated as one homogenous group. Because there were no significant differences between the different SGR-SE conditions at the posttests on any dependent measure ($ps > .10$), we treat them as one collapsed SGR-SE condition in the following analyses.

To investigate the hypothesis that working with the MGR-SE version of the tutor leads to higher learning gains than working

with the SGR-SE version on all measures of robust learning, we used the ANCOVA model to compute the main effect of number of graphical representations on students' test scores at the immediate and delayed posttests with pretest score as a covariate. The main effect of number of graphical representations was significant on reproduction with the number line, $F(1, 445) = 9.02, p < .01$, partial $\eta^2 = .02$, conceptual transfer, $F(1, 445) = 7.01, p < .01$, partial $\eta^2 = .02$, and marginally significant on procedural transfer, $F(1, 445) = 3.29, p < .10$, partial $\eta^2 = .01$. We did not find an interaction between number of graphical representations and posttest time (i.e., immediate and delayed posttest) for any dependent measure, $ps > .10$. A priori contrasts comparing the MGR-SE and SGR-SE conditions showed a significant advantage for the MGR-SE condition on reproduction with the number line both at the immediate posttest, $t(445) = 2.09, p < .05, d = 0.09$, and at the delayed posttest, $t(445) = 2.66, p < .01, d = 0.12$, as well as on conceptual transfer at the delayed posttest, $t(445) = 2.27, p < .05, d = 0.10$.

As an alternative test for our hypothesis, we used post hoc comparisons within ANOVA model to investigate learning gains by condition (see Table 3). Pairwise comparisons showed that students in the MGR-SE condition improved significantly from pretest to immediate posttest on reproduction with area models, $t(445) = 2.40, p < .05, d = 0.10$, reproduction with the number line, $t(445) = 3.16, p < .01, d = 0.14$, as well as from pretest to delayed posttest on reproduction with the number line, $t(445) = 3.80, p < .01, d = 0.17$, and conceptual transfer, $t(445) = 2.63, p < .05, d = 0.12$, and marginally significantly on reproduction with area models, $t(445) = 2.05, p < .10, d = 0.09$. Students in the SGR-SE condition showed marginal improvement from pretest to delayed posttest on conceptual transfer, $t(330) = 2.05, p < .10, d = 0.06$.

Discussion

Our results show that students significantly improved from pretest to posttest *only* when they worked with the MGR-SE version of the Fractions Tutor, but not if they worked with the SGR-SE versions. We found that, while students in the MGR-SE condition show significant and lasting learning gains on most dependent measures (Table 3, row "Post hoc comparisons of learning gains for MGR-SE condition"), the collapsed SGR-SE condition shows marginal learning gains only on conceptual transfer at the delayed posttest, but not on any other dependent measure (Table 3, row "Post hoc comparisons of learning gains for SGR-SE condition"). We did not find improvement on procedural transfer in either condition. Procedural transfer was assessed with comparison tasks that required students to convert given fractions to a common denominator, or to find benchmarks to compare the

Table 3
Overview of Results From ANOVAs on Pretest, Immediate Posttest, and Delayed Posttest From Experiment 2

Effect	Posttest time	Reproduction of knowledge						Transfer of knowledge		
		Area models		Number lines		Conceptual transfer		Procedural transfer		
		Direction	Significance	Direction	Significance	Direction	Significance	Direction	Significance	
Main effect of test time										
Post hoc comparisons of learning gains for MGR-SE condition	Immediate	Posttests > pretest	<i>p</i> < .05	Posttests > pretest	<i>p</i> < .01	Posttests > pretest	<i>p</i> < .01			<i>ns</i>
	Delayed	Posttest > pretest	<i>p</i> < .05	Posttest > pretest	<i>p</i> < .01	Posttest > pretest	<i>p</i> < .05			<i>ns</i>
Post hoc comparisons of learning gains for SGR-SE condition	Immediate	Posttest > pretest	<i>p</i> < .10	Posttest > pretest	<i>p</i> < .01	Posttest > pretest	<i>p</i> < .05			<i>ns</i>
	Delayed		<i>ns</i>		<i>ns</i>		<i>ns</i>			<i>ns</i>

Note. ANOVA = analysis of variance; MGR = multiple graphical representation; SE = self-explanation; SGR = single graphical representation.

Table 4
Overview of Results From ANCOVAs on Immediate Posttest and Delayed Posttest With Pretest as Covariate From Experiment 2

Effect	Reproduction of knowledge						Transfer of knowledge		
	Area models		Number lines		Conceptual transfer		Procedural transfer		
	Posttest time	Direction	Significance	Direction	Significance	Direction	Significance	Direction	Significance
Main effect of number of representations									
A priori comparisons for the effect of number of representations	Immediate	<i>ns</i>	MGR-SE > SGR-SE	<i>p</i> < .01	MGR-SE > SGR-SE	<i>p</i> < .01	MGR-SE > SGR-SE	<i>p</i> < .10	
	Delayed	<i>ns</i>	MGR-SE > SGR-SE	<i>p</i> < .05	MGR-SE > SGR-SE	<i>p</i> < .05	MGR-SE > SGR-SE	<i>ns</i>	<i>ns</i>

Note. ANCOVA = analysis of covariance; MGR = multiple graphical representation; SE = self-explanation; SGR = single graphical representation.

fractions to. Although the Fractions Tutor does not provide practice with these procedures, we expected that students would acquire knowledge about the relative size of fractions that they could use to solve the procedural transfer tasks. However, the results are not in line with this expectation. The procedural transfer tasks may have demanded too much of the students in our sample. In other words, we believe that the lack of learning gains on procedural transfer is due to a misalignment of this test scale and the Fractions Tutor. On the remaining test scales, the MGR-SE condition shows significant learning gains.

In line with our hypothesis that students working with multiple graphical representations would learn more, we found an advantage of the MGR-SE condition over the SGR-SE conditions on several dependent measures (see Table 4). Specifically, we found advantages of working with MGRs over working with an SGR on reproduction items that included number lines as well as on conceptual transfer, but not (as we had hypothesized) on reproduction with area models or on procedural transfer. This finding suggests that MGRs are particularly useful at promoting conceptual knowledge of fractions. As argued, different graphical representations provide different conceptual perspectives on fractions, which might encourage students to engage in deeper processing of fractions concepts and to form a more comprehensive mental model. Procedural knowledge, on the other hand, requires students to learn how to carry out algorithms. It appears that MGRs do not promote the acquisition of transferable procedural knowledge.

Further, we found that MGRs help learning of number lines, but not of area models. It is encouraging that MGRs promote learning about the number line because the number line is an important, central representational tool in mathematics. It can be used to connect fractions to real numbers and decimals, and it is a foundation for understanding coordinate systems in later algebra (NMAP, 2008; Siegler et al., 2010). On the other hand, area models may be more intuitive and familiar for students than number lines (e.g., because they build on students' real-world knowledge about sharing and division activities, see Mack, 1995). Furthermore, area models tend to be introduced earlier in fractions instruction than number lines. Perhaps the greater familiarity explains that the MGR-SE version of the Fractions Tutor did not help students perform better on reproduction with area models, compared to the SGR-SE conditions.

General Discussion

The goal of the experiments presented in this article was to investigate whether the well-established advantage of multiple external representations (i.e., text and graphic) generalizes to an advantage of multiple *graphical* representations over a single *graphical* representation when each is presented in addition to text and numbers. We focus on situations in which students encounter graphical representations one-at-a-time, hypothesizing that students will form a more accurate mental model of the domain knowledge by gradually refining it as they encounter different graphical representations that emphasize complementary conceptual aspects of fractions. This question is interesting from a practical standpoint because MGRs are typically used in realistic educational materials. This question is also interesting from a theoretical standpoint because existing theoretical frameworks for learning with multiple external representations do not make spe-

cific predictions as to whether MGRs are more effective than an SGR. Specifically, the CTML (Mayer, 2003, 2005) suggest that MGRs may be effective because they might enhance active integration and deeper conceptual processing than an SGR, or because they can yield more elaborate mental models of the domain content. Similarly, the ITPC (Schnottz & Bannert, 2003; Schnottz, 2005) suggests that MGRs might be more effective than an SGR because they enable students to form a more elaborate mental model of the domain. However, the ITPC also cautions that the potential advantage of MGRs needs to outweigh the costs associated with understanding each graphical representation. Investigating whether MGRs lead to better learning than an SGR is a step toward closing the gap between, on one hand, educational psychology research that has mostly focused on learning with multiple external representations and, on the other hand, common practice of using MGRs in instructional materials.

Our two experiments provide evidence that MGRs can lead to better learning than SGRs, when they are accompanied by self-explanation prompts. We attribute our finding to the complementary conceptual perspectives that MGRs provide on the learning content. As in many STEM domains, instruction on fraction uses different graphical representations with the goal to emphasize different conceptual aspects of the domain (Charalambous & Pitta-Pantazi, 2007). Only if students integrate these different conceptual views into one coherent mental model can they gain full conceptual understanding of fractions. Deep conceptual processing of complex learning material may be crucial to students' benefit from MGRs, in line with both the CTML and the ITPC. Our work extends these frameworks by showing that learning can be enhanced by integrating multiple graphical representations, presented consecutively across different problems.

Although this integration process is critical to students' benefit from multiple representations, students do not (often) engage in it spontaneously (Ainsworth et al., 2002; Yerushamly, 1991) and thus need to be supported in doing so. We implemented instructional support for this critical process in the form of self-explanation prompts that encourage students to make connections between each graphical representation and the key concepts of fractions they depict. In Experiment 1, we found that self-explanation prompts were in fact *necessary* for students in our experiment to benefit from MGRs: Only when provided with self-explanation prompts did we find an advantage of MGRs over SGRs. Although several studies have demonstrated that students benefit from self-explaining multirepresentational learning materials (e.g., Ainsworth & Loizou, 2003; Berthold et al., 2009; Zhang & Linn, 2011), Experiment 1 is, to the best of our knowledge, the first to systematically investigate the effects of self-explanation prompts while contrasting MGRs versus SGRs. It thus extends research that has investigated effects of self-explanation prompts on learning with multiple representations.

Given that MGRs were presented consecutively in our experiments, the self-explanation prompts likely stimulated gradual refinement of students' mental model of fractions, as they encountered additional conceptual aspects across a sequence of fractions problems. As mentioned, we chose to present graphical representations across consecutive problems because we consider this to be the next logical step in extending research on multiple external representations to multiple graphical representations, because, in our experiments, the number of direct connections between repre-

sentations was held constant across conditions. Furthermore, concurrent presentation of MGRs may place high demands on cognitive load. An interesting open question regards whether our findings generalize to other possible ways to present MGRs within problem sequences, for instance, when MGRs are presented concurrently, within the same problem. In light of our findings in Experiment 1, we expect that students' success in learning from MGRs will depend on their ability to relate each of them to key domain concepts. Based on prior research documenting that students tend not to spontaneously engage in such sense-making processes (Ainsworth et al., 2002; Yerushamly, 1991), we expect that even when MGRs are presented concurrently, students may need to be supported to engage in these processes. Yet presenting MGRs concurrently would allow students to make connections directly between conceptually corresponding elements of graphical representations. On the one hand, support for connection making between MGRs might further enhance students' benefits from MGRs because students can then directly compare the different conceptual aspects that each graphical representation emphasizes. On the other hand, it may be that this task is cognitively overwhelming, such that cognitive overload interferes with students' learning. It would be interesting to investigate whether direct support for connection making between graphical representations further enhances students' benefits from MGRs, and how this support should be designed such that potential negative effects due to high cognitive load can be prevented.

Our experiments showed different effects of MGRs on different types of knowledge. We found that MGRs promote learning of conceptual knowledge, and (in Experiment 2) that that MGRs help students learn the more difficult graphical representation. Since different graphical representations provide different conceptual perspectives on the abstract concept of fractions, they might enhance deeper processing of crucial concepts within the domain, leading to an advantage on conceptual knowledge. However, counter to our hypothesis, MGRs did not enhance procedural knowledge. In retrospect, we can see why MGRs might not be particularly helpful to students' learning of procedures performed on fractions. It may be that the ability to perform procedures on fractions is independent of the graphical representation used, such that performing these algorithms on different graphical representations does not enhance students' learning of procedural knowledge (e.g., fraction addition). There is a hint in the results from Experiment 1 to indicate that MGRs may enhance students' ability to transfer procedural knowledge to novel tasks, but overall, the evidence that MGRs enhance procedural transfer is rather weak and should be explored in future research. Given these considerations, we expect that our findings generalize to other domains that use MGRs to enhance students' conceptual knowledge by providing different conceptual perspectives on the domain, each instantiated by a particular graphical representation. It is likely that there are many such domains, including STEM domains.

While the effect sizes in Experiment 1 were considerable (ranging between $d = 0.44$ and $d = 0.99$), we found only small effect sizes in Experiment 2. We attribute the small effect sizes in Experiment 2 to the small learning gains; it may be that a difference of $d = 0.12$ between conditions when the learning gains are only $d = 0.17$ is meaningful. It may be that these effect sizes reflect the fact that the students in our studies already had done a considerable amount of fractions learning before the study started

(in fact, students in Experiment 2 came from a higher performing student population than those in Experiment 1). It may be, further, that these effect sizes reflect that fact that learning with MGRs is not without its cost (as pointed out by the theoretical frameworks). Not only must students become familiar with the different graphical representations, the sense-making processes that are required to take advantage of these representations to build richer, more integrated mental models may impose substantial cognitive load. It may be that additional interventions that support students in making direct connections between MGRs in a way that decreases cognitive load (e.g., using color coding to direct students' attention to relevant conceptual aspects) would further increase the effects of MGRs. Our results may not (yet) justify building up a practice of working with MGRs in a domain in which such a practice has not yet been established. On the other hand, the practical relevance of an intervention depends not only on effect sizes but also on the ease with which it is implemented. As discussed, it is common practice to use MGRs in many STEM domains. Our results provide support for that practice.

In sum, the work presented here demonstrates that students' learning can benefit from MGRs in a complex and challenging area of mathematics learning within the context of realistic educational settings. It extends the literature on learning with multiple external representations in several important ways. To the best of our knowledge, it is the first rigorous experimental investigation that compares learning with MGRs to learning with an SGR, each provided in addition to text and numbers. Across two classroom experiments, our results consistently demonstrate that students' robust learning of conceptual knowledge of fractions can be enhanced by providing them with MGRs, as long as students are prompted to self-explain the relation of each graphical representation to the key concepts it depicts.

References

Ainsworth, S. (2006). DeFT: A conceptual framework for considering learning with multiple representations. *Learning and Instruction*, 16, 183–198. doi:10.1016/j.learninstruc.2006.03.001

Ainsworth, S., Bibby, P., & Wood, D. (2002). Examining the effects of different multiple representational systems in learning primary mathematics. *Journal of the Learning Sciences*, 11, 25–61. doi:10.1207/S15327809JLS1101_2

Ainsworth, S., & Loizou, A. T. (2003). The effects of self-explaining when learning with text or diagrams. *Cognitive Science: A Multidisciplinary Journal*, 27, 669–681. doi:10.1016/S0364-0213(03)00033-8

Aleven, V., & Koedinger, K. R. (2002). An effective metacognitive strategy: Learning by doing and explaining with a computer-based Cognitive Tutor. *Cognitive Science: A Multidisciplinary Journal*, 26, 147–179. doi:10.1207/s15516709cog2602_1

Aleven, V., McLaren, B., Sewall, J., & Koedinger, K. (2009). A new paradigm for intelligent tutoring systems: Example-tracing tutors. *International Journal of Artificial Intelligence in Education*, 19, 105–154.

Arcavi, A. (2003). The role of visual representations in the learning of mathematics. *Educational Studies in Mathematics*, 52, 215–241. doi:10.1023/A:1024312321077

Atkinson, R. K., Renkl, A., & Merrill, M. M. (2003). Transitioning from studying examples to solving problems: Effects of self-explanation prompts and fading worked-out steps. *Journal of Educational Psychology*, 95, 774–783. doi:10.1037/0022-0663.95.4.774

Bennett, J. M. (2004). *Holt middle school math*. New York, NY: Holt Rinehart & Winston.

Berthold, K., Eysink, T. H. S., & Renkl, A. (2009). Assisting self-explanation prompts are more effective than open prompts when learning with multiple representations. *Instructional Science*, 37, 345–363. doi:10.1007/s11251-008-9051-z

Bodemer, D., Plötzner, R., Bruchmüller, K., & Häcker, S. (2005). Supporting learning with interactive multimedia through active integration of representations. *Instructional Science*, 33, 73–95. doi:10.1007/s11251-004-7685-z

Charalambous, C. Y., & Pitta-Pantazi, D. (2007). Drawing on a theoretical model to study students' understandings of fractions. *Educational Studies in Mathematics*, 64, 293–316. doi:10.1007/s10649-006-9036-2

Chi, M. T., Bassok, M., Lewis, M. W., Reimann, P., & Glaser, R. (1989). Self-explanations: How students study and use examples in learning to solve problems. *Cognitive Science: A Multidisciplinary Journal*, 13, 145–182. doi:10.1016/0364-0213(89)90002-5

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cook, M., Wiebe, E. N., & Carter, G. (2008). The influence of prior knowledge on viewing and interpreting graphics with macroscopic and molecular representations. *Science & Education*, 92, 848–867. doi:10.1002/sce.20262

Cramer, K., Wyberg, T., & Leavitt, S. (2008). The role of representations in fraction addition and subtraction. *Mathematics teaching in the middle school*, 13, 490–496.

de Jong, T., Ainsworth, S. E., Dobson, M., Van der Meij, J., Levonen, J., Reimann, P., . . . Swaak, J. (1998). Acquiring knowledge in science and mathematics: The use of multiple representations in technology-based learning environments. In M. W. Van Someren, W. Reimers, H. P. A. Boshuizen, & T. de Jong (Eds.), *Learning with multiple representations* (pp. 9–40). Bingley, England: Emerald.

Etel, A., Scheiter, K., & Schüller, A. (2013). How inspecting a picture affects processing of text. *Applied Cognitive Psychology*, 27, 451–461. doi:10.1002/acp.2922

Fitzgerald, F., Lappan, P., & Fey, J. T. (2004). *Connected mathematics (number and operations Grade 6, bits and pieces I understanding rational numbers)*. Upper Saddle River, NJ: Pearson Prentice Hall.

Gadgil, S., Nokes-Malach, T. J., & Chi, M. T. (2012). Effectiveness of holistic mental model confrontation in driving conceptual change. *Learning and Instruction*, 22, 47–61. doi:10.1016/j.learninstruc.2011.06.002

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science*, 7, 155–170. doi:10.1207/s15516709cog0702_3

Gobert, J. D., O'Dwyer, L., Horwitz, P., Buckley, B. C., Levy, S. T., & Wilensky, U. (2011). Examining the relationship between students' understanding of the nature of models and conceptual learning in biology, physics, and chemistry. *International Journal of Science Education*, 33, 653–684. doi:10.1080/09500691003720671

Gutwill, J. P., Frederiksen, J. R., & White, B. Y. (1999). Making their own connections: Students' understanding of multiple models in basic electricity. *Cognition and Instruction*, 17, 249–282. doi:10.1207/S1532690XCI1703_2

Hake, S. (2004). *Saxon math 5/6*. Norman, OK: Saxon.

Johnson, C. I., & Mayer, R. E. (2010). Applying the self-explanation principle to multimedia learning in a computer-based game-like environment. *Computers in Human Behavior*, 26, 1246–1252. doi:10.1016/j.chb.2010.03.025

Kafai, Y. B., Franke, M. L., Ching, C. C., & Shih, J. C. (1998). Game design as an interactive learning environment for fostering students' and teachers' mathematical inquiry. *International Journal of Computers for Mathematical Learning*, 3, 149–184. doi:10.1023/A:1009777905226

Koedinger, K. R., & Aleven, V. (2007). Exploring the assistance dilemma in experiments with cognitive tutors. *Educational Psychology Review*, 19, 239–264. doi:10.1007/s10648-007-9049-0

Koedinger, K. R., & Corbett, A. (2006). *Cognitive Tutors: Technology bringing learning sciences to the classroom*. New York, NY: Cambridge University Press.

Kordaki, M. (2010). A drawing and multi-representational computer environment for beginners' learning of programming using C: Design and pilot formative evaluation. *Computers & Education*, 54, 69–87. doi:10.1016/j.compedu.2009.07.012

Kozma, R., & Russell, J. (2005). Students becoming chemists: Developing representational competence. In J. K. Gilbert (Ed.), *Models and modeling in science education: Vol. 1. Visualization in science education* (pp. 121–145). Houten, the Netherlands: Springer Netherlands. doi:10.1007/1-4020-3613-2_8

Mack, N. (1995). Confounding whole-number and fraction concepts when building on informal knowledge. *Journal for Research in Mathematics Education*, 26, 422–441. doi:10.2307/749431

Mayer, R. E. (2003). The promise of multimedia learning: Using the same instructional design methods across different media. *Learning and Instruction*, 13, 125–139. doi:10.1016/S0959-4752(02)00016-6

Mayer, R. E. (2005). Cognitive theory of multimedia learning. In R. E. Mayer (Ed.), *The Cambridge handbook of multimedia learning* (pp. 31–48). Cambridge, England: Cambridge University Press. doi:10.1017/CBO9780511816819.004

Moss, J., & Case, R. (1999). Developing children's understanding of the rational numbers: A new model and an experimental curriculum. *Journal for Research in Mathematics Education*, 30, 122–147. doi:10.2307/749607

Nathan, M. J., Walkington, C., Srisurichan, R., & Alibali, M. W. (2011, June). *Modal engagements in precollege engineering: Tracking math and science concepts across symbols, sketches, software, silicone, and wood*. Paper presented at the American Society for Engineering Education, Vancouver, British Columbia, Canada. Available at <http://www.asee.org/public/conferences/1/papers/315/view>

National Mathematics Advisory Panel. (2008). *Foundations for success: Report of the National Mathematics Advisory Board Panel*. Washington, DC: U.S. Government Printing Office.

Pane, J. F., Griffin, B. A., McCaffrey, D. F., & Karam, R. (2013). Effectiveness of cognitive tutor algebra I at scale. *Educational Evaluation and Policy Analysis*, 36, 127–144. doi:10.3102/0162373713507480

Rau, M., Aleven, V., Rummel, N., & Rohrbach, S. (2013). Why interactive learning environments can have it all: Resolving design conflicts between competing stakeholder goals. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 109–118). New York, NY: Association for Computing Machinery.

Rau, M. A., Rummel, N., Aleven, V., Pacilio, L., & Tunc-Pekkan, Z. (2012). How to schedule multiple graphical representations? A classroom experiment with an intelligent tutoring system for fractions. In J. van Aalst, K. Thompson, M. J. Jacobson, & P. Reimann (Eds.), *The future of learning: Proceedings of the 10th International Conference of the Learning Sciences* (Vol. 1, pp. 64–71). International Society of Learning Sciences.

Scheiter, K., Gerjets, P., Huk, T., Imhof, B., & Kammerer, Y. (2009). The effects of realism in learning with dynamic visualizations. *Learning and Instruction*, 19, 481–494. doi:10.1016/j.learninstruc.2008.08.001

Schnotz, W. (2005). An integrated model of text and picture comprehension. In R. E. Mayer (Ed.), *The Cambridge handbook of multimedia learning* (pp. 49–70). Cambridge, England: Cambridge University Press. doi:10.1017/CBO9780511816819.005

Schnotz, W., & Bannert, M. (2003). Construction and interference in learning from multiple representation. *Learning and Instruction*, 13, 141–156. doi:10.1016/S0959-4752(02)00017-8

Siegler, R. S., Carpenter, T., Fennell, F., Geary, D., Lewis, J., Okamoto, Y., . . . Wray, J. (2010). *Developing effective fractions instruction: A practice guide*. Washington, DC: U.S. Department of Education.

Sweller, J., van Merriënboer, J. J. G., & Paas, F. G. W. C. (1998). Cognitive architecture and instructional design. *Educational Psychology Review*, 10, 251–296. doi:10.1023/A:1022193728205

van der Meij, J., & de Jong, T. (2006). Supporting students' learning with multiple representations in a dynamic simulation-based learning environment. *Learning and Instruction*, 16, 199–212. doi:10.1016/j.learninstruc.2006.03.007

van der Meij, J., & de Jong, T. (2011). The effects of directive self-explanation prompts to support active processing of multiple representations in a simulation-based learning environment. *Journal of Computer Assisted Learning*, 27, 411–423. doi:10.1111/j.1365-2729.2011.00411.x

Van Someren, M. W., Boshuizen, H. P. A., & de Jong, T. (1998). Multiple representations in human reasoning. In M. W. Van Someren, H. P. A. Boshuizen, & T. de Jong (Eds.), *Learning with multiple representations* (pp. 1–5). Oxford, England: Pergamon.

Wylie, R., & Chi, M. T. H. (in press). The self-explanation principle in multimedia learning. In R. Mayer (Ed.), *The Cambridge handbook of multimedia learning* (2nd ed.). Cambridge, England: Cambridge University Press.

Yerushamly, M. (1991). Student perceptions of aspects of algebraic function using multiple representation software. *Journal of Computer Assisted Learning*, 7, 42–57. doi:10.1111/j.1365-2729.1991.tb00223

Zhang, Z. H., & Linn, M. C. (2011). Can generating representations enhance learning with dynamic visualizations? *Journal of Research in Science Teaching*, 48, 1177–1198. doi:10.1002/tea.20443

(Appendix follows)

Appendix

Supplementary Information on Materials and Results

Table A1
Estimated Marginal Means (and Standard Deviations) for Experiment 1

Posttest time	Variable	SGR-noSE		SGR-SE		MGR-noSE		MGR-SE	
		<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
Immediate posttest	Conceptual reproduction	2.45	0.67	2.32	0.96	1.92	1.12	2.75	0.64
	Procedural reproduction	2.95	0.62	2.79	0.62	2.68	0.84	2.95	0.58
	Conceptual transfer	1.60	0.64	1.60	0.92	1.65	1.01	1.71	1.14
	Procedural transfer	2.23	1.26	1.65	1.27	1.27	1.21	2.36	0.84
Delayed posttest	Conceptual reproduction	1.98	1.10	1.97	1.09	1.42	1.03	2.40	0.70
	Procedural reproduction	2.55	0.73	2.57	0.66	2.31	0.88	2.73	0.48
	Conceptual transfer	2.30	0.87	2.21	0.99	2.06	1.03	2.66	0.58
	Procedural transfer	2.39	0.96	2.11	1.02	1.94	1.14	2.35	0.77

Note. SGR = single graphical representation; SE = self-explanation; MGR = multiple graphical representation. The maximum score was 3 for all knowledge types.

Table A2
Estimated Marginal Means (and Standard Deviations) for Experiment 2

Time	Variable	SGR-SE		MGR-SE	
		<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
Pretest	Reproduction with number line	0.43	0.03	0.45	0.03
	Reproduction with area models	0.61	0.03	0.59	0.03
	Conceptual transfer	0.68	0.03	0.73	0.03
	Procedural transfer	0.49	0.04	0.58	0.04
Immediate posttest	Reproduction with number line	0.50	0.03	0.60	0.03
	Reproduction with area models	0.64	0.03	0.70	0.03
	Transfer conceptual	0.74	0.03	0.79	0.03
	Transfer procedural	0.46	0.04	0.55	0.04
Delayed posttest	Reproduction with number line	0.51	0.03	0.63	0.03
	Reproduction with area models	0.65	0.03	0.68	0.03
	Transfer conceptual	0.74	0.03	0.84	0.03
	Transfer procedural	0.52	0.04	0.52	0.04

Note. SGR = single graphical representation; SE = self-explanation; MGR = multiple graphical representation. The maximum score was 1 for all knowledge types.

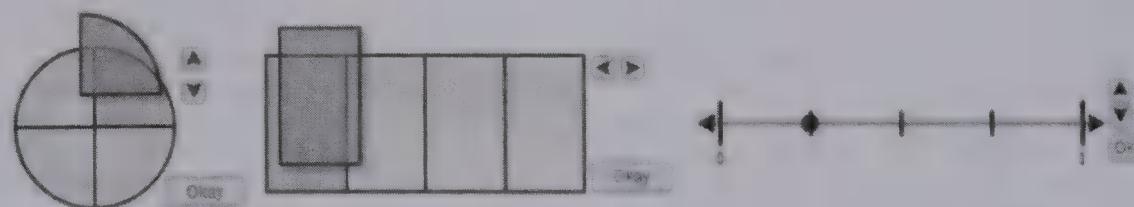


Figure A1. Interactive representations used in fractions tutor: circle, rectangle, and number line.

(Appendix continues)

This is the unit.

Let's make $\frac{5}{9}$.

First, partition the circle into equal sections.

Next, drag one section into the white circle diagram.

This is $\frac{1}{9}$ of

Circle A:

To show $\frac{5}{9}$ you need to make copies of $\frac{1}{9}$.

This is the unit.

Let's make $\frac{1}{8}$.

First, partition the circle into equal sections.

Next, drag one section into the white circle diagram.

This is $\frac{1}{8}$ of

Circle B:

To show $\frac{5}{9}$ you need to make copies of $\frac{1}{8}$.

Circle A has equal sections. Circle B has equal sections.

The sections in circle A are than the sections in circle B, because the circle A has sections.

Circle A has the same number of colored sections as circle B.

Therefore, $\frac{5}{9}$ is $\frac{5}{8}$.

smaller than
larger than
equal to

Figure A2. Making a circle given a symbolic fraction, combined with prompts to compare the two fractions. Reflection prompts are implemented with drop-down menus shown at the bottom.

This is the unit.

Please make $\frac{2}{7}$.

Partition the rectangle into equal sections.

Drag one section into the white rectangle.

This is $\frac{1}{7}$ of

Rectangle A:

To show $\frac{2}{7}$ you need to make copies of $\frac{1}{7}$.

This is the unit.

Please make $\frac{2}{5}$.

Partition the rectangle into equal sections.

Drag one section into the white rectangle.

This is $\frac{1}{5}$ of

Rectangle B:

To show $\frac{2}{5}$ you need to make copies of $\frac{1}{5}$.

Rectangle A has total sections. Rectangle B has total sections.

The sections in rectangle A are than the sections in rectangle B, because the rectangle A has sections.

Rectangle A has the same number of colored sections as rectangle B.

Therefore, $\frac{2}{7}$ is than $\frac{2}{5}$.

Congratulations! You're done!

Done

Figure A3. Making a rectangle given a symbolic fraction given a symbolic fraction, combined with prompts to compare the two fractions. Reflection prompts are implemented with drop-down menus shown at the bottom.

(Appendix continues)

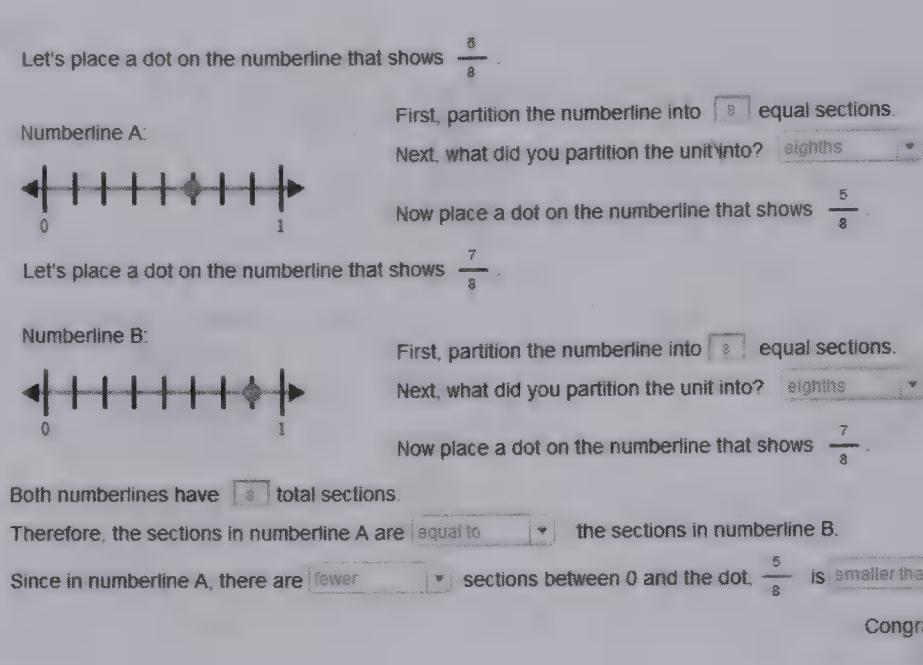


Figure A4. Showing a fraction on the number line given a symbolic fraction, combined with prompts to compare the two fractions. Reflection prompts are implemented with drop-down menus shown at the bottom.

Received September 28, 2012
 Revision received May 14, 2014
 Accepted May 19, 2014 ■

ORDER FORM

Start my 2015 subscription to the *Journal of Educational Psychology*® ISSN: 0022-0663

____	\$99.00	APA MEMBER/AFFILIATE	_____
____	\$229.00	INDIVIDUAL NONMEMBER	_____
____	\$821.00	INSTITUTION	_____
Sales Tax: 5.75% in DC and 6% in MD			
TOTAL AMOUNT DUE		\$	_____

Subscription orders must be prepaid. Subscriptions are on a calendar year basis only. Allow 4-6 weeks for delivery of the first issue. Call for international subscription rates.



SEND THIS ORDER FORM TO
 American Psychological Association
 Subscriptions
 750 First Street, NE
 Washington, DC 20002-4242

Call 800-374-2721 or 202-336-5600
 Fax 202-336-5568 TDD/TTY 202-336-6123
 For subscription information,
 e-mail: subscriptions@apa.org

Check enclosed (make payable to APA)

Charge my: Visa MasterCard American Express

Cardholder Name _____

Card No. _____ Exp. Date _____

Signature (Required for Charge)

Billing Address

Street _____

City _____ State _____ Zip _____

Daytime Phone _____

E-mail _____

Mail To

Name _____

Address _____

City _____ State _____ Zip _____

APA Member # _____

An Imagination Effect in Learning From Scientific Text

Claudia Leopold
University of Muenster

Richard E. Mayer
University of California, Santa Barbara

Asking students to imagine the spatial arrangement of the elements in a scientific text constitutes a learning strategy intended to foster deep processing of the instructional material. Two experiments investigated the effects of mental imagery prompts on learning from scientific text. Students read a computer-based text on the human respiratory system (control group), read while being asked to form an image corresponding to each of 9 paragraphs (imagery group), or read while being asked to form an image and with seeing an onscreen drawing before each paragraph (picture-before-imagery group) or after each paragraph (picture-after-imagery group). Imagery prompts facilitated transfer and retention performance compared to a control group on an immediate test (Experiment 1: $d = 1.30$ on transfer, $d = 0.74$ on retention) and on a delayed test (Experiment 2: $d = 0.86$ on transfer, $d = 0.98$ on retention), but the added drawings had no additional effect. The findings support the imagination principle, which states that people learn more deeply when prompted to form images depicting the spatial arrangement of what they are reading.

Keywords: imagination, imagery, multimedia learning, learning strategy

Consider a text that explains how the respiratory system works, such as shown in Appendix A. What can be done to help students learn more deeply so that they are better able to answer transfer questions based on the lesson? One approach is to add graphics, such as a graphic for each paragraph that depicts the structure or functioning of a portion of the respiratory system as described in the paragraph. The rationale for this approach comes from research on the *multimedia principle*, which has shown that students learn more deeply from words and graphics than from words alone (Butcher, 2014; Mayer, 2009). For example, Mayer (2009) reported that across more than a dozen experimental comparisons, students performed better on a transfer test after reading a scientific passage accompanied by corresponding graphics (e.g., drawings or animation) than without graphics, yielding a median effect size greater than 1.

The explanation for the multimedia principle is that students given words and graphics are more likely to engage in appropriate cognitive processing during learning, including selecting corresponding information in the text and graphics, organizing this information into corresponding cognitive representations, and integrating the verbal and pictorial representations with each other and with relevant prior knowledge (Mayer, 2009). In his dual coding theory, Paivio (1986, 2007) pointed to the positive cogni-

tive consequences that occur when learners make *referential connections* between words and images during learning. Similarly, the cognitive theory of multimedia learning posits that building connections between corresponding verbal and pictorial representations is a central process in meaningful learning, as indicated by superior transfer performance.

The present study takes the multimedia principle one step further by asking whether students can learn more deeply by imagining the spatial arrangement of elements described in a scientific text about how the respiratory system works. We call this process *seeing with the mind's eye* because the students engage in multimedia learning by imagining internal graphics rather than viewing external graphics. The goal of the present study is to determine the cognitive consequences of asking students to imagine graphics that depict the structure and functioning of the respiratory system being described in the text. Overall, we aim to test what can be called the *imagination principle*, which posits that students learn more deeply from an explanatory scientific text when they are asked to form mental images corresponding to the structures and processes described in the text. The present study is motivated by the relative lack of research on the imagination principle as an aid to understanding explanatory text (Dunlosky, Rawson, Marsh, Nathan, & Willingham, 2013).

Literature Review

This study on the imagination principle is motivated in part by recognition that the potentially powerful role of mental imagery in human learning, memory, and cognition has been examined across a variety of research literatures including spatial cognition, verbal learning, memory mnemonics, and educational psychology.

In spatial cognition, a number of studies have examined basic characteristics and functions of mental imagery (Farah, 1984; Ganis, Thompson, & Kosslyn, 2004; Johansson, Holsanova, & Holmqvist, 2006; Shepard & Cooper, 1982). An important finding

This article was published Online First June 16, 2014.

Claudia Leopold, Institute for Psychology in Education, University of Muenster; Richard E. Mayer, Department of Psychological and Brain Sciences, University of California, Santa Barbara.

This research was supported by a fellowship of the German Research Foundation granted to Claudia Leopold.

Correspondence concerning this article should be addressed to Claudia Leopold, Institute for Psychology in Education, University of Muenster, 48149 Muenster, Germany. E-mail: claudia.leopold@psy.uni-muenster.de

from these studies is that the cognitive processing of imagined representations follows similar mechanisms as the cognitive processing of perceived representations (Borst & Kosslyn, 2012; Finke, 1985; Kosslyn, Thompson, & Ganis, 2006). Thus, there is a functional equivalence between visual mental imagery and visual perception. This functional equivalence refers, for example, to how people inspect and rotate mental images (Kosslyn, Ball, & Reiser, 1978; Shepard & Metzler, 1971) and indicates that the underlying representations for these imagery processes are depictive and spatial in nature.

In verbal learning, studies conducted by Paivio (1986) and his coworkers showed that mental imagery enhances recall performance on basic memory tasks such as remembering word lists. An important finding is the concreteness effect, in which people remember lists of concrete words or sentences better than lists of abstract words or sentences (see Paivio, 1965, 1969, for a review; Sadoski, Goetz, & Fritz, 1993). Sadoski, Goetz, and Rodriguez (2000) and Goolsby and Sadoski (2013) reported similar findings for concrete versus abstract texts. Paivio explained these results with the idea that concrete words and sentences evoked imagery processes that aided their recall. This interpretation is supported by the results of Sadoski and Quast (1990), who found close relations between students' imagery ratings of text passages and their long-term recall ($r = .40$). Furthermore, Paivio and his colleagues reported that an instruction to imagine lists of concrete nouns versus an instruction to pronounce these nouns improved recall probability by about 50% (Paivio, 1975; Paivio & Csapo, 1973). Both of these findings can be explained by Paivio's dual coding theory, which posits that adding a nonverbal imaginal code to a verbal code serves as a supplementary route for facilitating recall (see also Sadoski & Paivio, 2013).

In memory mnemonics (e.g., keyword method, method of loci), the idea of dual coding is applied to facilitate recall of vocabulary items (Atkinson, 1975; Raugh & Atkinson, 1975), technical terminology and foreign words (Carney & Levin, 1998; Jones, Levin, Levin, & Beitzel, 2000), and facts (Brigham & Brigham, 1998; Levin, Morrison, McGivern, Mastropieri, & Scruggs, 1986; McCormick, Levin, & Valkenaar, 1990). These mnemonic techniques have in common that they rely on imagery processes in order to establish, for example, referential connections between a vocabulary item (e.g., the German word *Fenster* = *window*) and an acoustic associative that is similar in sound (e.g., *faint*). An example is a mental image of a person standing before a window and suddenly fainting so that he or she is falling into the window. According to Paivio (1986), these kinds of images help students to build connections between verbal and nonverbal (imagery) representations.

In educational psychology, two branches of research on mental imagery can be identified, focusing on the role of imagination in learning procedures and in learning facts. In the first research branch, focusing on memory for procedures, researchers have established an imagination effect when students imagine their actions as they learn a procedural task. For example, students were asked to imagine the steps of a procedure for how to construct formulae in a spreadsheet application (Cooper, Tindall-Ford, Chandler, & Sweller, 2001), how to apply geometry rules (Ginns, Chandler, & Sweller, 2003), how to find a route in a bus timetable (Leahy & Sweller, 2005), or how to use a temperature line graph (Leahy & Sweller, 2005). Overall, instructions to form mental

images in these experiments facilitated learning the various procedural tasks when the students had sufficient prerequisite schemata about the task. The authors explained this effect with the idea that imagination requires learners to automatize the procedures similar to mental practice in perceptual-motor tasks frequently investigated in sports psychology (Driskell, Copper, & Moran, 1994). As this type of imagery strategy is focused on facilitating automation of procedures, it corresponds to an imagery rehearsal strategy.

In the second branch of research, focusing on memory for facts, research established an imagination effect when students were asked to imagine pictures in their mind corresponding to facts in a narrative. For example, in a classic study, Pressley (1976) taught elementary school children in a 20-minute training how to form mental images and asked them afterward to read a 950-word story with the instruction to make up pictures in their head as they read. Students in the control group received a control training in which they were asked to do whatever they could in order to remember the story. The results showed that students in the imagery group remembered more facts about the story than the students in the control group. Similar results were reported by Gambrell and Jawitz (1993) with fourth-grade students, by Kulhavy and Swenson (1975) with sixth-grade children, and by Giesen and Peeck (1984) and Rasco, Tennyson, and Boutwell (1975, Experiment 1) with college students.

In contrast, Anderson and Kulhavy (1972) and Rasco et al. (1975, Experiment 2) showed no effect of an imagery instruction on text recall, but Anderson and Kulhavy found that students who actually reported using the imagery strategy performed better in a recall test than students who reported not using mental imagery. Thus, it seems important to provide clear and specific imagery instructions and to check whether the students really follow these instructions.

In general, these results suggest that imaging a picture while reading a story is a powerful strategy for fostering recall of facts. Furthermore, the results of Rasco et al. (1975) and Gambrell and Jawitz (1993) showed that there was no difference between students who were asked to create mental pictures and students who were provided with external pictures. Thus, the same underlying processes may apply to both internally constructed images and externally presented images.

The present study extends the study of imagination effects involving procedural knowledge (by imagining carrying out steps in a procedure) and factual knowledge (by imagining mental pictures about a story) to an imagination effect involving conceptual knowledge (by forming an image of the spatial structure of a scientific system). Investigating the effects of imagination on conceptual knowledge is relevant because scientific texts often remain challenging for students (Best, Rowe, Ozuru, & McNaught, 2005; Graesser, 2007). VanLehn and colleagues (2007, Experiment 2), for example, found no learning gains from reading passages from a physics textbook compared to students who read nothing at all but just took the test. Graesser (2007) pointed out that scientific or technical text is a challenge because students often lack relevant background knowledge and adequate reading strategies directed at facilitating deep comprehension. To our knowledge, the present study is the first to examine whether imagination strategies can apply to the educationally relevant

domain of learning how a conceptual system works, using problem-solving transfer as a dependent measure.

Theory and Predictions

According to the cognitive theory of multimedia learning, meaningful learning (as measured by transfer test performance) occurs when learners engage in appropriate cognitive processing during learning, including selecting relevant verbal and visual material from the lesson, mentally organizing it into verbal and pictorial representations, and integrating the representations with each other and with relevant prior knowledge activated from long-term memory. Prompts to imagine the spatial arrangement of elements in scientific text are intended to prime these processes in the same way that providing well-designed graphics creates a multimedia effect (Mayer, 2009).

One process that is crucial in both, in processing text with mental imagery and in processing text with corresponding pictures, is the integration of words and images, that is, creating referential connections between words and corresponding images. In mental imagery, the process of creating referential connections is essential because mental imagery cannot be applied without the learner drawing connections between words or phrases and their corresponding images (Sadoski & Paivio, 2013). This integration of words and images is a key ingredient in *generative processing* in the cognitive theory of multimedia learning; therefore, mental imagery can be considered a generative learning strategy (Mayer, 2009). Similarly, in learning with text and pictures, referential connections between words and corresponding pictures are crucial for facilitating deeper understanding of the text content, that is, to transfer knowledge to new problems (Kester, Kirschner, & van Merriënboer, 2005; Mayer, 2009; Mayer, Steinhoff, Bower, & Mars, 1995). If creating referential connections is crucial for developing a deep understanding of the learning materials and if we take into account that referential connections are a key component of the imagination process, then imagery activities should improve transfer and retention test performance.

Furthermore, when students build images of the spatial relationships that are expressed in the text, these mental imagery activities can facilitate mental model building (Johnson-Laird, 1983). This spatial form of mental imagery promotes an internal representation that preserves topological relations between elements of a system and therefore structural equivalence with the referential system (Denis, 2008; Denis & Cocude, 1989). On the basis of this internal representation, students can derive structural knowledge about the major components of the system as well as dynamic knowledge about how the system works (Mayer & Gallini, 1990). This is consistent with the view that learners can build runnable mental models of a dynamic system (Hegarty, 2004). Mental imagery can therefore be called a model-focused strategy that should affect the students' ability to transfer their knowledge to new problems.

To our knowledge, there are no studies that directly test the effects of mental imagery in learning from explanatory scientific text on transfer performance. Leutner, Leopold, and Sumfleth (2009) found an interaction between drawing instruction and imagery instruction in terms of imagery instruction facilitating comprehension in the absence of drawing instruction. However, the comprehension test required students to draw text-based inferences but did not include transfer questions. The results of Leopold,

Sumfleth, and Leutner (2013) showed that student's self-reported mental imagery activities partly mediated the students' spatial representations about the text content, which in turn mediated transfer performance. These studies support the idea that imagery affected the students' spatial representations and deeper understanding, although this idea was not directly tested.

One problem that may affect the effectiveness of mental imagery concerns the quality of students' created images. Denis and Cocude (1992) observed that students had difficulties in constructing accurate mental images from a text that described the spatial outline of a fictive island. Denis (2008) related these difficulties to two processes—construction and review processes. Constructing mental images is a sequential process in which students generate images and step by step add one image to the other. By contrast, reviewing mental images involves the activation of the whole image so that the image can be used for manipulation or comparison tasks. Although constructing and reviewing are dynamic processes that are intertwined, constructing is usually more important in the beginning of a learning phase, while reviewing is more important at the end of a learning phase (Denis, 2008). To support students in constructing mental images, we presented external pictures before the students read and imagined each text paragraph. In this sense, the picture provides a scaffold for the imagery process (Eitel, Scheiter, Schüller, Nyström, & Holmqvist, 2013). To support students in reviewing and uploading their mental images, we presented external pictures after the students read and imagined each text paragraph. The picture provided external feedback for their mental image.

The theoretical rationale for studying the imagination principle is the same as for the multimedia principle, that is, both principles are based on the idea that deeper learning occurs when learners engage in the act of building connections between corresponding words and pictures that describe how a system works. In the case of the multimedia principle, the pictures are provided by the instructor, but in the case of the imagination principle, the pictures are imagined by the learner (with guiding instructions). This integration of words and graphics is called *generative processing* in the cognitive theory of multimedia learning and is posited to lead to meaningful learning outcomes.

In the present study, students read an explanatory scientific text on how the human respiratory system works (as shown in Appendix A) either with or without prompts to imagine (as shown in Appendix B). In addition, for some learners, pictures were provided as instructional support for the imagery process by presenting a picture before or after each text paragraph. Based on the cognitive theory of multimedia learning, we predicted that students who were asked to imagine corresponding graphics as they read an explanatory science text would score higher on subsequent transfer tests than students who simply read the text (Prediction 1). We also expected that students who were provided with external pictures to support the imagery process would score higher on transfer tests than students who simply imagined the text on their own (Prediction 2).

Secondary predictions were that students who were asked to imagine would also show superior performance on retention of the key steps in the explanation and on drawing the key steps in the explanation as compared to students who simply read (Predictions 3 and 5). Finally, we expected that students who were provided with external pictures to support the imagery process would score

higher on retention and drawing tests than students who simply imagined the text on their own (Predictions 4 and 6). In addition, as a preliminary step, we tested whether the treatment groups differed in time on task, self-reported motivation, perceived difficulty, and mental effort.

Experiment 1

Experiment 1 tested these six predictions on an immediate test.

Method

Participants and design. The participants were 85 college students recruited from the psychology subject pool of the University of California, Santa Barbara. Their mean age was 19.09 years ($SD = 1.14$), and the percentage of female students was 64.3%. They scored low on a survey of prior knowledge ($M = 3.28$, $SD = 2.07$, based on a 13-point measure), and their mean score on a 10-point test of spatial ability was 4.16 ($SD = 3.31$). The study was based on a between-subjects design with four levels of imagery instruction (imagery group, picture-before-imagery group, picture-after-imagery group, and control group). Twenty students served in the imagery group, 22 in the picture-before-imagery group, 20 in the picture-after-imagery group, and 23 in the control group.

Materials. The learning materials were computer based and consisted of four versions of a lesson on how the human respiratory system works adapted from a shorter lesson used by Mayer and Sims (1994). The text contained 786 words and consisted of an introduction and nine paragraphs. We computed a readability score using the Flesch-Kincaid grade level formula as an indicator of text difficulty (Kincaid, Fishburne, Rogers, & Chissom, 1975). The readability score of the text was 9.9, which indicates that the text was appropriate for students from Grades 10 and higher and was thus appropriate for college students.

The same text was used in all four versions and is reproduced in Appendix A. In all four versions, each paragraph was presented on a separate screen along with the following headings: (a) Structure of the Nervous System, (b) Steps in the Nervous System to Control Breathing, (c) Structure of the Thoracic Cavity, (d) Structure of the Airway System, (e) Process of Inhaling, (f) Structure of the Exchange System, (g) Structure of the Circulatory System, (h) Process of Exchanging, and (i) Process of Exhaling. Students clicked on a *next* button in order to move from one paragraph to the next paragraph. The presentations were developed using Macromedia Authorware 7.0.

The control version of the lesson included just the text paragraphs with the *next* button presented below each paragraph, as exemplified in Figure 1. The imagery version was identical to the control version except that a specific imagery instruction was added to the right of each paragraph, for example, "Please imagine the steps in the nervous system when the brain sends a signal to the diaphragm and rib muscles." Figure 2 shows a screenshot from the imagery version of the lesson. The imagination instructions for each of the nine paragraphs are listed in Appendix B. The picture-before-imagery version of the lesson was identical to the imagery version except that a drawing was presented before each paragraph. The drawing depicted the content of the following paragraph. In order to move from the picture to the corresponding

Structure of the Nervous System

The respiratory center is located in the rear, bottom part of the brain, near the back of the neck. The respiratory center of the brain is connected to a pathway of nerves that leads down from the spinal cord to connect with muscles controlling the diaphragm and the rib cage.

next

Figure 1. A screenshot of the program presented to the control group. See the online article for the color version of this figure.

paragraph, the students clicked on the *next* button. In order to move from this paragraph to the next picture, they clicked on the *next* button again, and so on. The students only saw either the picture or the paragraph, never both of them at the same time. Figure 3 shows a screenshot of a picture shown before a paragraph in the lesson. The picture-after-imagery version was identical to the picture-before-imagery version except that the corresponding picture was presented after the students had read and imagined the respective paragraph with the instruction: "Please compare this with your mental picture." Thus, students in the picture-after-imagery group first saw the paragraph with the imagery instruction, then clicked the *next* button and saw the corresponding picture. When they clicked on the *next* button again, they moved on to the next paragraph, and so on.

The testing materials consisted of a retention test, a transfer test, a drawing test, a paper-folding test, and a questionnaire. The testing materials were printed on 8.5-in. \times 11-in. sheets of paper.

The retention test contained the following instruction at the top of the sheet: "Using what you learned in the session, please write an explanation of how the human respiratory system works." For scoring the student's explanations, we divided the text into 35 idea units based on the paragraphs about the process of respiration and 41 idea units based on the paragraphs about the structure of the

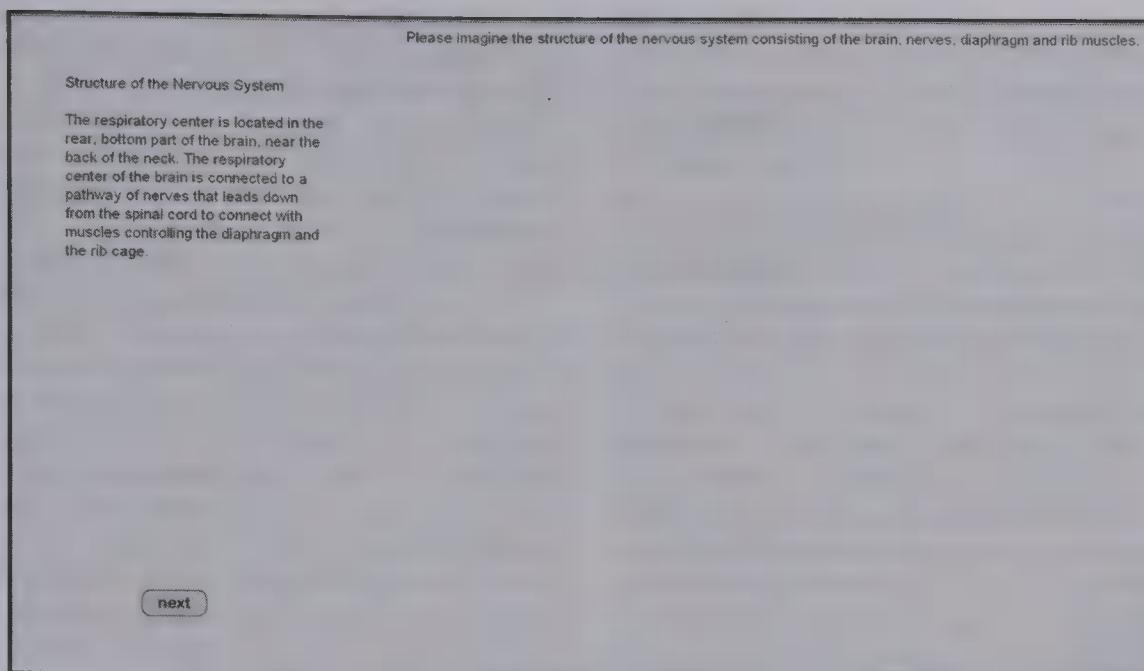


Figure 2. A screenshot of the program presented to the imagery group. See the online article for the color version of this figure.

respiratory system. The headings for the paragraphs on the process of respiration were (a) Steps in the Nervous System to Control Breathing, (b) Process of Inhaling, (c) Process of Exchanging, and (d) Process of Exhaling. We computed a process-retention score for each student by counting the number of ideas (out of 35) that the student included in his or her explanation. One point was given for correctly stating each of the 35 idea units, for example, "brain detects the need for oxygen," "brain sends out a signal to inhale," "signal moves to muscles controlling the diaphragm or rib cage," "the diaphragm contracts downward," and "the rib cage moves slightly outward." The headings of the paragraphs about the structure of the respiratory system were (a) Structure of the Nervous

System, (b) Structure of the Thoracic Cavity, (c) Structure of the Airway System, (d) Structure of the Exchange System, and (e) Structure of the Circulatory System. We computed a structure-retention score for each student by counting the number of ideas (out of 41) that the student included in his or her explanation. One point was given for correctly stating each of the 41 idea units, for example, "respiratory center is located in rear part of brain," "from brain nerves lead down the spinal cord," "nerves lead to muscles of the diaphragm and rib cage," "the thoracic cavity contains the lungs," "the thoracic cavity is surrounded by ribs," "ribs can move inward or outward," and "the diaphragm is on the bottom of the thoracic cavity." The participant did not have to show the exact

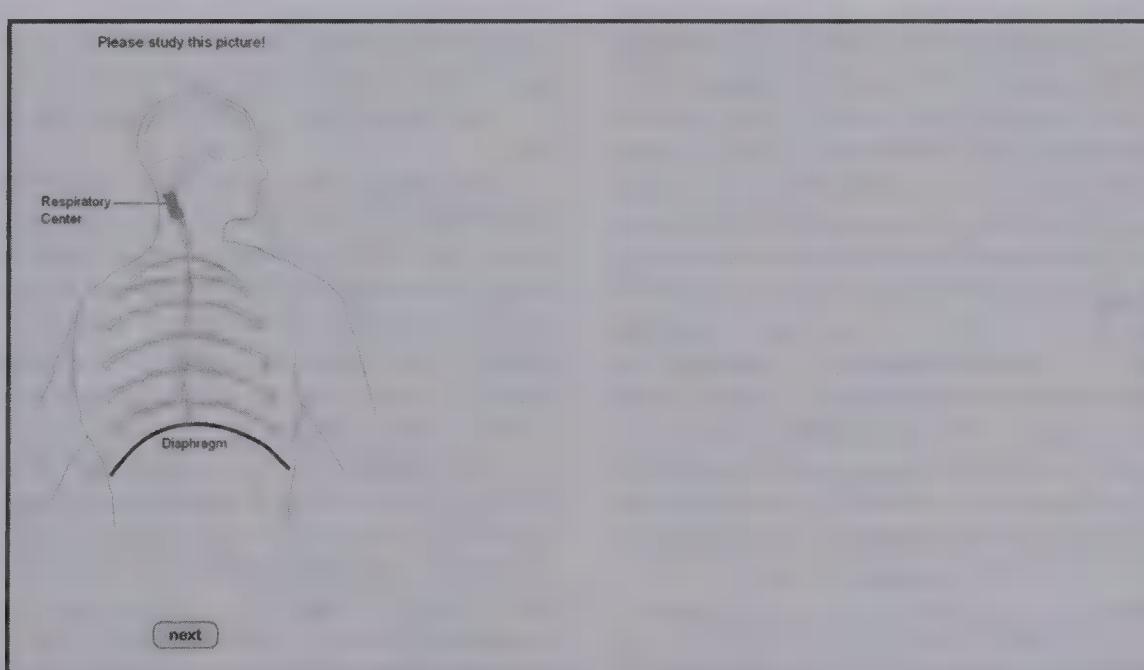


Figure 3. A screenshot of a picture shown to the picture-before-imagery group. See the online article for the color version of this figure.

wording or correct spelling to receive credit for an idea unit but had to express the correct idea. All scoring was done by consensus between two raters who were blind to the participants' group. Separate scores for process and structure were computed in order to determine whether the effects of imagining helped students to visualize the static structure of the respiratory system and the dynamic functioning of the respiratory system. Interrater reliabilities based on 25% of the data were $r = .93$ for the process-retention test and $r = .79$ for the structure-retention test.

The transfer test consisted of five sheets, each containing a question that required the students to apply their knowledge to new problems, such as "Although there is oxygen in the lungs, the cells in the body do not get enough oxygen to make energy. What could have caused this problem?" or "Suppose you are a scientist who is trying to improve the human respiratory system for people who climb high mountains (where less oxygen is in the air). What could be done to make the human respiratory system more effective for mountain climbing?" We computed a transfer score for each participant by counting the number of acceptable answers across the five transfer questions (Cronbach's $\alpha = .66$). The reliability of $\alpha = .66$ is acceptable but a bit lower than expected, which may depend on the fact that only five transfer questions were used. Interrater reliability based on 25% of the data was $r = .98$. Acceptable answers for the first question were air cannot get into the air sacs, capillaries cannot pick up oxygen from the air sacs, the arteries are blocked, veins do not take away carbon dioxide, the connection between lungs and heart is blocked, blockage in the bronchioles, heart does not beat regularly, cells cannot absorb oxygen, and so on. Acceptable answers for the second question were expand the rib muscles, expand the diaphragm, expand the lung's capacity, change the regulation of the brain's system, absorb more oxygen with every breath, make the exchange system more effective, add air sacs in the lungs, add something that can bind more oxygen in the blood, add more channels of capillaries or arteries, and so on. The transfer test was intended to assess the learner's depth of understanding of the material and therefore is the primary learning outcome measure in this study.

The drawing test contained the following two instructions, each typed on a separate sheet: "Please draw a picture of the exchange system and label the different parts." "Please draw a picture of the respiratory system when the person inhales and label the different parts." These instructions referred to the representation of key components of the respiratory system explained in the text and their spatial relations. The students were informed that sketching the important components and their interrelations would be sufficient rather than drawing aesthetically appealing pictures. The accuracy of the exchange drawing was assessed using a checklist that consisted of seven criteria based on seven components of the exchange system and their spatial location, that is, the lungs, the alveoli, the capillaries, oxygen, carbon dioxide, arteries, and veins. The accuracy of the inhaling drawing was assessed using a checklist based on four criteria, that is, the windpipe-to-lung connection, expansion of lungs, flattening of the diaphragm, and expansion of ribs. An accurate drawing of each component was given 2 points, a partly accurate component was given 1 point, and an unacceptable drawing of a component received 0 points. For example when a student's drawing showed the diaphragm beneath the lungs and the student had indicated (by arrows or by labels) that the shape of the diaphragm was flattened, 2 points were given. When it was not

obvious that the diaphragm was flattened, 1 point was given. When the diaphragm was drawn in the incorrect shape or when the diaphragm was not mentioned at all, 0 points were given. The maximal number of points was 22 for the two drawings (Cronbach's $\alpha = .72$). Two raters scored 25% of the students' drawings with an interrater reliability of $r = .95$. The drawing test was intended to assess the quality of the student's self-generated visual-spatial representations of the respiratory system.

The paper-folding test consisted of a sheet with 10 problems taken from Ekstrom, French, Harman, and Dermen (1976). Each item required imaging a paper being folded, punched with a hole, and reopened. One point was given for each correct answer, and one point was subtracted for each wrong answer, with a total possible score of 10. The paper-folding test was intended to measure an aspect of spatial ability that has been found to be related to mental imagery (Denis, 2008).

The first sheet of the questionnaire consisted of seven self-report scales on the students' motivation, their effort/difficulty, and strategy use. The motivation questions were "I enjoyed learning from this lesson," "I would like to learn from more lessons like this," and "Please rate how appealing this lesson was for you." Each was accompanied by a 5-point scale ranging from *strongly agree* to *strongly disagree* (Cronbach's $\alpha = .84$). The effort and difficulty questions were "Please rate how difficult this lesson was for you" (with a 5-point scale from *very easy* to *very difficult*) and "Please rate how much effort you exerted in learning this lesson" (with a 5-point scale ranging from *very low* to *very high*). The final two questions were "Please rate your spatial ability" (with a 5-point scale from *very low* to *very high*) and "I prefer to learn visually" (with a 5-point scale from *strongly agree* to *strongly disagree*).

The second sheet consisted of a four-item self-report questionnaire on mental imagery and six distractor items. The four items were translated from a version used by Leopold et al. (2013) and were used to check whether students followed the instructions during the study phase (with Cronbach's $\alpha = .85$). All self-report scales were 5-point scales ranging from 1 (*strongly disagree*) to 5 (*strongly agree*). The items were "I mentally imagined how the processes described in the text work," "I formed mental pictures about the text content in order to understand it," "I created mental pictures about the text content," and "I tried to understand the structures and processes of the respiratory system by mental imagery."

The final sheet of the questionnaire included questions concerning the students' age, gender, and prior knowledge about the human body. Their knowledge of the human body was measured by using a 13-item checklist. Students were asked to "place a check mark next to the things that apply to you," based on the following list: "I have participated in science programs or fairs," "Biology was my favorite subject in high school," "I sometimes watch science documentaries about anatomy in my free time," "I can name most of the components of the human heart from memory," "I have taken a course in human anatomy or physiology," "I attended a course on cardiopulmonary resuscitation (CPR) training," "I can explain what pulmonary embolism means," "I sometimes find myself on the Internet looking up biology related topics," "I know the difference between venous and arterial blood," "I have watched an educational video on how the respiratory system works," "I talked to a doctor about how the process of respiration works," "I know the definition of the terms 'dia-

stolic' and 'systolic,' and "I took advanced biology classes in high school (AP, IB, Honors), etc." A prior knowledge score was computed by tallying the number of items checked on the checklist, yielding a maximum score of 13.

Procedure. Students were randomly assigned to the experimental groups and were tested in groups of one to three per session. Each student was seated at an individual cubicle in front of a computer. First, participants signed an informed-consent sheet. Then, the experimenter presented oral instructions stating that the students would receive a lesson on the human respiratory system that would comprise an introduction and nine paragraphs. They were told to go at their own pace and that when they were finished, the experimenter would have some questions for them to answer. Students pressed the space bar and then saw a page that welcomed them and thanked them for their participation in the experiment. This page repeated the instructions that they would read an introduction and nine paragraphs on the respiratory system and should click on the *next* button in order to move through the presentation. Furthermore, students in the three imagery groups were informed that first they would receive a short pretraining in how to use their mental imagination. This pretraining consisted of examples of how to imagine text content based on two text paragraphs about the global warming effect. The text in each paragraph was presented sentence by sentence. After reading the first sentence, students pressed the space bar and then saw how a possible image of that sentence would look. When they pressed the space bar again, the next sentence was presented, and after pressing the space bar again, the image was adapted to the content of the new sentence, and so on. On average students spent 95.87 s ($SD = 39.62$) on this pretraining of how to form mental images.

After the pretraining, the imagery groups were presented with a page that told them that now they would receive the lesson on the human respiratory system and that they were asked to imagine the text as a graphic. The students in the picture-before-imagery group were informed that in order to help them to imagine the information, they would be shown a drawing before each paragraph. The students in the picture-after-imagery group were informed that in order to help them, they would be shown a drawing after they had imagined the paragraph so that they could compare their mental picture with the presented drawing. The students then studied the respective version of the presentation on the respiratory system according to their experimental group.

When the presentation was finished, the experimenter presented instructions for the paper-folding test, and then, the students were given the paper-folding test for 3 minutes. After 3 minutes, the experimenter collected the paper-folding test and distributed the retention sheet, which asked the students to write an explanation of how the human respiratory system works. After 5 minutes, the retention sheet was collected, and the transfer sheets were presented one at a time. Students were given 2.5 minutes for each transfer question. Each transfer sheet was collected by the experimenter before the next sheet was presented. Afterward, the two sheets of the drawing test were distributed one at a time. Students were asked (a) to draw a picture of the exchange system and label the different parts and (b) to draw a picture of the respiratory system when the person inhales and to label the different parts. Students were told that their picture did not have to be beautiful but could be simple and should depict the important parts. Two and a half minutes were given for each of the drawing tasks. Then, the experimenter distributed each sheet of the questionnaire,

with instructions to answer as honestly as possible and to complete the questionnaire at their own rate. Upon completion, participants were thanked and excused. We followed guidelines for ethical treatment of human subjects.

Results and Discussion

We computed analyses of variance to test overall differences among the experimental groups. To test predictions requiring a comparison of the three treatment means with the mean of the control group, we used Dunnett tests. Dunnett tests control the familywise Type I error rate at 5% and is at the same time a powerful test specifically designed to compare each treatment with a control (Klockars & Sax, 1986; Sheskin, 2011). To test predictions requiring comparisons between the picture-imagery groups and the imagery group, we used Bonferroni's correction as it is suitable for a small number of comparisons and can be transferred to nonorthogonal comparisons.

Before testing the effects of the imagery instructions on the dependent variables, we examined whether the four treatment groups were equivalent on basic characteristics and whether the groups followed their particular instruction.

Are the groups equivalent on basic characteristics? Analyses of variance (with $p < .05$) showed that the groups did not differ on prior knowledge, spatial ability, age, their self-rated spatial ability, and their preference for learning visually. A chi-square analysis revealed there was a difference in the proportion of males and females.¹

Did the students follow the instructions? In the self-report questionnaire, we asked the students whether they really had imagined the text content. The top line of Table 1 shows the mean imagery score (and standard deviation) for each group, with higher scores indicating higher degrees of imagery during learning. We used these data as a manipulation check. An analysis of variance revealed a significant effect of treatment, $F(3, 81) = 4.25$, $MSE = .47$, $p = .008$, $\eta^2 = .14$. Dunnett tests showed that the picture-before-imagery group ($p = .053$), the imagery group ($p = .007$), and the picture-after-imagery group ($p = .012$) each reported more mental imagery activity during learning than the control group did. We take this as evidence that the imagination prompts were successful in promoting imagery during learning from the science text. Table 1 shows that the control group spontaneously reported imagining to a substantial degree, which can be explained by the fact that the text used concrete language, which has been shown to evoke mental images (see the review of Sadoski & Paivio, 2013). Our results indicate that explicit strategy instruction enhanced this effect.

Does imagery instruction facilitate transfer performance (Prediction 1)? The primary research question addressed in this study concerns whether students learn more deeply from a science text when they are prompted to form mental images of the respiratory system and how it works as they read. Based on the imagination hypothesis, students who form mental images of the

¹ We computed analyses of covariance with gender as a covariate for all of the performance measures. The results did not change except for the main effect of treatment on process-retention scores. Although gender was not a significant predictor, $F(3, 80) = 1.54$, $p = .218$, the main effect of treatment did not remain significant ($p = .109$).

Table 1

Experiment 1: Means, Standard Deviations, and Effect Sizes for Self-Reported Mental Imagery, Motivation, Perceived Difficulty, and Mental Effort

Self-report ratings	Experimental group											
	Control		Imagery			Picture-imagery			Imagery-picture			
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>d</i>	<i>M</i>	<i>SD</i>	<i>d</i>	<i>M</i>	<i>SD</i>	<i>d</i>	
Mental imagery	3.68	0.89	4.34 ^a	0.55	0.92	4.17 ^a	0.47	0.72	4.30 ^a	0.74	0.75	
Motivation	3.03	0.92	3.22	0.82	0.21	3.70 ^a	0.62	0.87	3.53	0.95	0.53	
Perceived difficulty	3.30	0.93	3.00	0.92	-0.33	2.95	0.95	-0.37	2.85	0.88	-0.50	
Mental effort	3.00	0.80	3.10	0.72	0.13	3.05	0.90	0.06	2.85	1.09	-0.16	

^a Indicates significant difference from the control group.

system described in a science text should understand the material more deeply—through building connections between corresponding words and images—and therefore perform better on tests of problem-solving transfer. The top row of Table 2 presents the means (and standard deviations) of each group on the transfer test. An analysis of variance conducted on these data demonstrated a significant effect of treatment, $F(3, 81) = 4.70$, $MSE = 15.30$, $p = .004$, $\eta^2 = .15$. Dunnett tests showed that the imagery group outperformed the control group ($p = .001$), and there was no significant difference between the picture-before-imagery group and the control group ($p = .098$) or the picture-after-imagery group and the control group ($p = .180$). In line with our prediction, the superiority of the imagery group over the control group is new evidence for an imagination effect in which students learn more deeply when they are asked to form mental images of an explanatory scientific text. This finding is a primary contribution of this study.

Do the imagery groups that received drawings show better transfer performance than the pure imagery group (Prediction 2)? The means in the top row of Table 2 seem to indicate that the groups that received drawings and imagery prompts (i.e., picture-before-imagery and picture-after-imagery groups) showed lower scores in their transfer performance than the imagery group did. Planned comparisons revealed that the mean transfer score of the imagery group did not differ significantly from the picture-before-imagery group, $t(81) = 1.66$, $p = .200$, or from the picture-after-imagery group, $t(81) = 1.86$, $p = .134$. As can be seen and contrary to our predictions, providing drawings of the human respiratory system (in the picture-before-imagery group or

picture-after-imagery group) does not add to the effectiveness of imagination prompts, perhaps because the students did not have to work as hard to mentally construct their illustrations.

Does imagery instruction facilitate retention performance (Prediction3)? The foregoing analysis shows an imagination effect in which asking students to mentally create drawings for a science text results in improvements in transfer test performance, indicating deeper learning. A second research question concerns whether imagination prompts also help students better remember the structure and process of the system described in a scientific text. The second row of Table 2 shows each group's mean retention score (and standard deviation) for text describing the process of respiration, whereas the third row shows the mean retention scores (and standard deviations) for text describing the structure of the respiratory system. According to the imagination hypothesis, instructions to form mental images for the content of a scientific text should result in better memory for the process and structure of the system described in the text.

With regard to process-retention scores shown in the second line of Table 2, an analysis of variance revealed a significant overall effect of treatment, $F(3, 81) = 3.35$, $MSE = 13.06$, $p = .023$, $\eta^2 = .11$. To examine whether the three experimental groups who received an imagery instruction performed better than the control group did, we computed Dunnett tests. The results showed that the picture-before-imagery group ($p = .042$), the imagery group ($p = .052$), and the picture-after-imagery group ($p = .02$) each performed better than the control group did. With regard to structure-retention scores shown in the third line of Table 2, an analysis of variance revealed a significant overall effect of treatment, $F(3,$

Table 2

Experiment 1: Means, Standard Deviations, and Effect Sizes for the Learning Outcome Variables

Learning outcome scores	Experimental group											
	Control		Imagery			Picture-imagery			Imagery-picture			
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>d</i>	<i>M</i>	<i>SD</i>	<i>d</i>	<i>M</i>	<i>SD</i>	<i>d</i>	
Transfer test	6.13	3.14	10.60 ^a	3.72	1.30	8.59	4.67	0.63	8.30	3.99	0.61	
Process retention	6.57	3.63	9.20 ^a	3.50	0.74	9.23 ^a	4.04	0.69	9.60 ^a	3.19	0.89	
Structure retention	2.17	1.80	4.15 ^a	2.76	0.87	1.77	2.18	-0.20	2.15	1.73	-0.01	
Drawing test	8.09	4.23	11.30 ^a	3.77	0.80	13.06 ^a	4.31	1.16	11.80 ^a	4.76	0.83	
Study time (in seconds)	284.93	69.93	351.62	103.98	0.77	402.15 ^a	170.30	0.98	406.91 ^a	148.65	1.12	

^a Indicates significant difference from the control group.

$81) = 5.14$, $MSE = 4.60$, $p = .003$, $\eta^2 = .16$. Dunnett tests showed that the imagery group significantly outperformed the control group ($p = .011$), and there was no significant difference between the picture-before-imagery group and the control group ($p = .873$) or the picture-after-imagery group and the control group ($p = .999$). The mean scores in the structure-retention test were quite low and indicate a floor effect. The lower scores can be explained by the fact that the students were asked to write an explanation "of how the human respiratory system works." This task did not require the students to write down structure information but focused on process information.

Overall, the reported results provide additional support for the imagination hypothesis, in which students learn better when they form mental images about the structure and process of the system described in a science text as they read. One limitation of the results is that with gender as a covariate, the overall effect on process-retention scores did not remain significant.

Do the imagery groups who received drawings show better retention performance than the imagery group (Prediction 4)? We compared the picture-before-imagery group and the picture-after-imagery group with the imagery group, respectively. With regard to the process-retention score, neither the picture-before-imagery group nor the picture-after-imagery group significantly differed from the pure imagery group, $t(81) < 1$. With regard to the structure-retention score, the imagery group outperformed the picture-before-imagery group as well as the picture-after-imagery group, $t(81) = 3.59$, $p = .002$, and $t(81) = 2.95$, $p = .008$, respectively. Overall, contrary to our predictions, there is no indication that adding pictures enhances the effectiveness of imagination prompts.

Does imagery instruction facilitate drawing performance (Prediction 5)? The imagination hypothesis predicts that asking students to form images during learning will improve their performance on a drawing test. The fourth line in Table 2 shows the mean drawing score (and standard deviation) for each group. An analysis of variance conducted on the data summarized in the fourth line of Table 2 demonstrated a significant effect of treatment, $F(3, 81) = 5.48$, $MSE = 18.35$, $p = .002$, $\eta^2 = .17$. Consistent with predictions, Dunnett tests showed that the picture-before-imagery group ($p = .001$), imagery group ($p = .044$), and picture-after-imagery group ($p = .016$) performed better than the control group did.

Do the imagery groups that received pictures show better drawing performance than the imagery group (Prediction 6)? Neither the picture-before-imagery group nor the picture-after-imagery group performed significantly better than the imagery group, $t(81) = 1.32$, $p = .382$, and $t(81) < 1$, respectively. Apparently, actually seeing a picture and simply imaging a picture based on a science text produced equivalent improvements on a subsequent drawing test. Contrary to our predictions but comparable to the results of the transfer and retention tests, adding pictures did not increase drawing performance.

Do the groups differ in time on task? The mean study times and standard deviations for each group are shown in Table 2. There was a significant difference among the treatment groups in study time, $F(3, 81) = 4.31$, $p = .007$, $\eta^2 = .14$. Dunnett tests showed that the picture-before-imagery group ($p = .009$) and the picture-after-imagery group ($p = .008$) spent more time with the presentation than the control group did, but the imagery group did not

differ significantly from the control group ($p = .226$). To take into account the difference in study time, we computed analyses of covariance (ANCOVAs) with study time as a covariate and the performance measures as dependent variables. The conclusions remain the same when we run an ANCOVA with time as the covariate. The results do not differ from the former analysis for the transfer score, $F(3, 80) = 5.08$, $MSE = 15.27$, $p = .003$, $\eta^2 = .16$; for the structure-retention score, $F(3, 80) = 5.28$, $MSE = 4.61$, $p = .002$, $\eta^2 = .17$; and for the drawing score, $F(3, 80) = 5.06$, $MSE = 18.54$, $p = .003$, $\eta^2 = .17$, except that the overall effect of the treatment on the process-retention score did not remain significant, $F(3, 80) = 2.65$, $p = .055$. The effect of the covariate time was in none of the analyses significant ($p > .149$). There was also no significant interaction between the treatment and time on any of the performance measures ($p > .200$).

Do the students differ in their motivation, perceived difficulty, and mental effort scores? Table 1 summarizes the mean motivation rating and mental effort scores (and standard deviations) for each group. With regard to motivation, there was a significant effect for treatment, $F(3, 81) = 2.88$, $MSE = .70$, $p = .041$, $\eta^2 = .10$. Dunnett tests indicated that the picture-before-imagery group reported more enjoyment with the lesson than the control group did ($p = .024$), but the picture-after-imagery group and the imagery group did not differ from the control group ($p = .131$ and $p = .809$, respectively). With regard to the perceived difficulty of the lesson and the mental effort invested in studying the lesson, there were no differences among the experimental groups, $Fs(3, 81) < 1$. Overall, there is no strong evidence that students who received imagery prompts reported more difficulty or effort than the control group. Only the picture-before-imagery group reported more motivation than the control group did. However, self-report measures may not be the best way to assess these factors as they can be influenced by a tendency to choose socially approved behaviors, student ability, instruction, context of assessment, and so on (e.g., Kruger & Dunning, 1999). Behavioral measures such as the students' persistence in studying further text passages, physiological measures, or dual task paradigms might provide more reliable data (Brünken, Seufert, & Paas, 2010; DeLeeuw & Mayer, 2008).

How do the four learning outcome measures relate to one another? Table 3 shows a correlation matrix for the four learning outcome measures, with significant correlations indicated in bold font. As can be seen, the transfer score correlates significantly with each of the other three scores, the drawing score correlates significantly with two other scores, the process-retention score correlates significantly with two other scores, and the structure-

Table 3
Correlations Among Transfer, Process-Retention, Structure-Retention, and Drawing Scores

Learning outcome scores	1	2	3	4
1. Transfer	—	.57	.32	.52
2. Process retention		—	.06	.53
3. Structure retention			—	.17
4. Drawing				—

Note. Significant correlations are indicated in bold.

retention score correlates significantly with one other score. Overall, the transfer score appears to be the most inclusive, suggesting that it may be the best measure of deep learning. Consistent with results of other studies (Leopold et al., 2013; Schwamborn, Mayer, Thillmann, Leopold, & Leutner, 2010), there is a high correlation between the drawing scores and the transfer and process-retention scores ($r = .52$, $r = .53$, respectively), indicating close connections between the quality of the spatial representation of the respiratory system and measures of deep learning.

Does drawing performance mediate the effect of imagination activity on transfer performance? The results reported above showed that the imagery group performed better on tests of problem-solving transfer than the control group. We expected that the effect of condition (control vs. imagery) on transfer performance would be mediated by the quality of the students' internal spatial representations of the respiration process assessed by their drawing performance. This mediation hypothesis is based on the idea that mental imagery instruction promotes an internal spatial representation that preserves structural equivalence with the referential system and therefore facilitates transfer. Following the procedure proposed by Baron and Kenny (1986), we performed simple and multiple regression analyses. First, we computed the direct effect of the independent variable condition (code 1 = control, code 2 = mental imagery) on the dependent variable (transfer performance): $\beta = .56$, $p < .001$ (see Figure 4). Second, we tested whether the independent variable (condition) affects the mediating variable (spatial representation): $\beta = .38$, $p = .013$; third, we tested whether the mediating variable (spatial representation) affects transfer performance: $\beta = .52$, $p < .001$. Multiple regression analysis revealed that the direct effect of condition on transfer performance was reduced when the effect of the mediating variable was controlled: $\beta = .38$, $p = .003$. To test whether the indirect effect, that is, the path from the independent variable *condition* via the mediating variable *spatial representation*, on transfer test performance is significant, we conducted the Sobel test (Sobel, 1982; see also MacKinnon, Lockwood, Hoffman, West, & Sheets, 2002). The indirect effect ($\beta = .38 \times .52 = .20$) was significant ($z = 2.16$, $p = .031$). Thus, condition (i.e., imagery instructions) influenced transfer performance by affecting the students' spatial representations, which in turn affected their transfer performance. These results support partial mediation of the effect of condition on transfer performance via the quality of the students' spatial representations of the respiration process.

Experiment 2

In Experiment 1, the main result was that mental imagery prompts facilitate transfer performance. In Experiment 2, we focused on examining whether the imagination effect is stable over a time delay. Therefore, we included only an imagery group and a

control group. According to multimedia theory and dual coding theory, when learners build connections between text and a mental image, they construct a deeper learning outcome with more retrieval routes, which should enhance performance over a time delay. Previous research on learning strategies further suggests that a particularly useful measure of the effectiveness of learning strategies involves performance on a delayed test of retention or transfer (Dunlosky et al., 2013), so in the interests of consistency, we sought to determine whether the findings of Experiment 1 could be replicated after a 2-day delay.

Method

Participants and design. The participants were 48 college students recruited from the paid psychology subject pool at the University of California, Santa Barbara. Their mean age was 19.73 years ($SD = 1.38$), and the percentage of female students was 77.1%. They scored low on a survey of prior knowledge ($M = 3.04$, $SD = 2.62$, based on a 13-point measure), and their mean score on a 10-point test of spatial ability was 5.42 ($SD = 3.59$). The study was based on a between-subjects design with two of the experimental conditions used in Experiment 1: imagery group and control group. Twenty-three students served in the imagery group, and 25 served in the control group. These conditions were identical to the ones used in Experiment 1 except that we did not immediately test their retention, transfer, and drawing performance but rather tested them after a delay of 2 days. Two students did not return for the second part of the study, so we excluded their data.

Materials. The learning and testing materials were identical to the ones used in Experiment 1 except that two experimental conditions instead of four were implemented. The reliability (Cronbach's alpha) of the scales was very similar to the ones reported in Experiment 1: $\alpha = .72$ for the transfer test, $\alpha = .75$ for the drawing test, $\alpha = .91$ for self-reported imagery, and $\alpha = .83$ for self-reported motivation.

Procedure. The procedure was identical to that used in Experiment 1 except that the study consisted of two parts. In the first part, the students studied the presentation on how the human respiratory system works and filled in the paper-folding test. Then, the students were dismissed and asked to come back after 2 days. Students were told not to study anything about the human respiratory system during the 2-day delay. In the second part of the study, the students completed the retention test, the transfer test, the drawing test, and the questionnaire as in Experiment 1. The tests were scored with the same procedures used in Experiment 1.

Results and Discussion

Are the groups equivalent on basic characteristics? Analyses of variance or chi-square tests (with $p < .05$) showed that the groups did not differ on prior knowledge score, spatial ability, self-rated spatial ability, preference for learning visually, mean age, or proportion of males and females.

Did the students follow the instructions? The top line of Table 4 shows the mean imagery score (and standard deviation) for both groups. The analysis of the students' answers on the questionnaire reveals that the students in the imagery group reported more mental imagery activity than the students in the control group did, $t(46) = 3.55$, $p = .001$, $d = 1.08$. We take this as evidence that

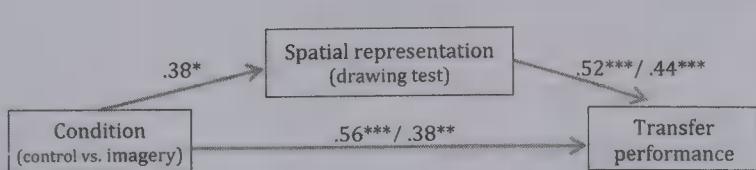


Figure 4. Mediation model in Experiment 1. * $p < .05$. ** $p < .01$. *** $p < .001$. See the online article for the color version of this figure.

Table 4

Experiment 2: Means, Standard Deviations, and Effect Sizes for Self-Reported Mental Imagery, Motivation, Perceived Difficulty, and Mental Effort

Self-report ratings	Experimental group				
	Control		Imagery		
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>d</i>
Mental imagery	3.37	1.11	4.29 ^a	0.60	1.08
Motivation	2.55	0.96	3.14 ^a	0.84	0.66
Perceived difficulty	3.32	1.18	3.52	0.85	0.20
Mental effort	2.75	0.99	3.22	1.00	0.47

^a Indicates significant difference from the control group.

the imagery group followed the imagery prompts during learning from the science text.

Does imagery instruction facilitate transfer performance (Prediction 1)? We hypothesized that students who form mental images of the system described in the science text should understand the material more deeply—and therefore perform better on tests of problem-solving transfer. The top row of Table 5 presents the means (and standard deviations) of the two groups on the transfer test. A *t* test revealed a significant difference among the groups, $t(46) = 2.93$, $p = .005$, $d = .86$, with the imagery group scoring higher than the control group. Thus, the imagery prompts facilitated problem-solving transfer compared to a control condition, even after a delay of 2 days.

Does imagery instruction facilitate retention performance (Prediction 3)? The second and third rows of Table 5 show the mean retention scores (and standard deviations) of the two groups. There was a significant effect of treatment on the process-retention score, $t(46) = 3.40$, $p = .001$, $d = .98$, and a nonsignificant marginal effect on the structure-retention score, $t(46) = 1.73$, $p = .092$, $d = .50$, with the imagery group scoring higher than the control group. Thus, imagery prompts facilitated retention of process information compared to a control condition even after a delay of 2 days.

Does imagery instruction facilitate drawing performance (Prediction 5)? The fourth line in Table 5 shows the mean drawing score (and standard deviation) for each group. There was a significant effect of treatment on the drawing score, $t(46) = 3.27$,

$p = .002$, $d = .95$, with the imagery group performing better than the control group.

Do the groups differ in time on task? Table 5 shows the mean study times and standard deviations for the two groups. There was a significant difference between the groups in study time, $t(46) = 2.99$, $p = .004$, $d = .89$, in which the imagery group spent longer with the presentation than the control group did. Therefore, we included study time as a covariate and computed ANCOVAs with the performance measures as dependent variables. For the transfer score, $F(1, 45) = 5.83$, $MSE = 18.11$, $p = .020$; process-retention score, $F(1, 45) = 9.39$, $MSE = 14.77$, $p = .004$; and drawing score, $F(1, 45) = 7.20$, $p = .010$, the results did not change; for the structure-retention score, there no longer was a marginally significant effect when study time was included as a covariate, $F(1, 45) = 2.64$, $p = .111$. The effect of the covariate time was in none of the analyses significant. There was also no significant interaction between the treatment and time on any of the performance measures (all F s < 1).

Do the students differ in their motivation and mental effort scores? Table 4 shows the mean ratings (and standard deviations) on motivation, perceived difficulty, and mental effort for the two groups. With regard to motivation, there was a significant difference between the groups, $t(46) = 2.29$, $p = .026$, $d = .66$, in which the imagery group reported more enjoyment with the lesson than the control group did. This result is in line with Sadoski and Quast (1990; see also Sadoski & Paivio, 2013), who reported associations between affective factors and mental imagery activity.

With regard to perceived difficulty and mental effort, there were no significant differences between the groups, $t(46) < 1$, and $t(45) = 1.61$, $p = .114$. These findings are consistent with Experiment 1.

How do the four learning outcome measures relate to one another? Table 6 shows a correlation matrix for the four learning outcome measures, with significant correlations indicated in bold font. Similarly to Experiment 1, the transfer score correlates significantly with each of the other three scores. This also applies for the drawing score, while the process-retention and structure-retention scores each correlate significantly with two other scores. There are strong correlations between the drawing score and the transfer score, as was found in Experiment 1. Overall, these results are very consistent with those of Experiment 1.

Does drawing performance mediate the effect of imagination activity on transfer performance? As in Experiment 1, we expected that the effect of condition (control vs. imagery) on transfer performance would be mediated by the quality of the students' internal spatial representations of the respiration process.

Table 5

Experiment 2: Means, Standard Deviations, and Effect Sizes for the Learning Outcome Variables

Learning outcome scores	Experimental group				
	Control		Imagery		
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>d</i>
Transfer test	6.64	3.38	10.22 ^a	4.99	0.86
Process retention	5.40	3.29	9.13 ^a	4.29	0.98
Structure retention	1.52	2.12	2.65	2.42	0.50
Drawing test	7.36	4.53	11.39 ^a	3.95	0.95
Study time (in seconds)	236.40	64.19	319.97 ^a	122.58	0.89

^a Indicates significant difference from the control group.

Table 6

Experiment 2: Correlations Among Transfer, Process-Retention, Structure-Retention, and Drawing Scores

Learning outcome scores	1	2	3	4
1. Transfer	—	.63	.40	.59
2. Process retention	—	—	.12	.53
3. Structure retention	—	—	—	.51
4. Drawing	—	—	—	—

Note. Significant correlations are indicated in bold.

Simple regression analyses demonstrated that condition (independent variable) significantly affected transfer performance ($\beta = .40$, $p < .003$) and the mediating variable spatial representation ($\beta = .44$, $p = .002$; see Figure 5). Also, the mediating variable spatial representation predicted transfer performance ($\beta = .59$, $p < .001$). Furthermore, multiple regression analysis revealed that the direct effect of condition on transfer performance was no longer significant when the effect of the mediating variable was controlled ($\beta = .17$, $p = .199$), whereas the effect of the mediating variable spatial representation was significant ($\beta = .52$, $p < .001$). These results indicate full mediation of the effect of condition on transfer performance via students' spatial representations.

General Discussion

Empirical Contributions

This study's primary empirical contribution is that prompts and instruction to imagine the process of respiration and the structure of the respiratory system while reading an explanatory text on how human respiration works facilitate transfer and retention performance on immediate and delayed tests. These results extend previous findings concerning rote memory of words or facts by demonstrating that mental imagery can also affect deeper understanding of explanatory text, as shown in superior transfer performance.

A secondary finding is that adding external pictures—presented either before or after imagining the relevant text paragraph—did not add any benefits beyond simply imagining. For example, on the transfer test, the imagery group outperformed the control group, but the picture-before-imagery and picture-after-imagery groups did not. On the structure-retention test, the imagery group outperformed the control group and outperformed the picture-before-imagery group and the picture-after-imagery group. Overall, the results of Experiment 1 indicate the pictures did not enhance the imagery process as was proposed but in some cases actually weakened it—and this pattern was similar for both of the imagery-and-picture groups, providing evidence for the consistency of the results.

How can these results be explained? The text consisted of nine paragraphs, and each was accompanied by a picture. The students of the picture-before- and picture-after-imagery groups may have relied on the external presentation of the picture rather than on investing effort in creating their own internal picture of the text content. The students of the picture-before imagery group may have focused on the external presentation of the picture rather than on creating an internal picture of the text content. Similarly, the students in the picture-after-imagery group knew that a picture was

being presented after each paragraph. In anticipating that a picture would be shown, they may have decided not to put too much effort in the imagination process but rather rely on the external picture. In both cases, processing the external pictures would draw resources away from the mental imagery process. Students might have primarily focused on processing features of the pictures that were not useful in building a dynamic mental representation. Processing the external pictures could also have imposed high cognitive load on the learner because the learner had to mentally hold the pictorial representation in his or her working memory in order to integrate it with the textual input (van Merriënboer & Ayres, 2005). Mental imagery without external pictures, however, seems to foster deeper processing of the text content, presumably because the students connected words and their corresponding images and relied on these images when constructing a coherent runnable model of the respiratory system. A similar result was reported by Schworm and Renkl (2006), who found that self-explanation prompts were less effective when instructor explanations were available than when they were not available, presumably because the learners relied upon the instructor explanations rather than investing effort in self-explanations. More research is necessary to disentangle the effects of external representations, such as pictures, and the strategic processes of the learner, such as mental imagery.

Third, the positive effect of mental imagination not only was found on an immediate test but was replicated on a delayed test that was administered 2 days after the study phase. This result is consistent with the findings of Sadoski and Quast (1990), who found close relations between mental imagery ratings and text recall after a delay of 16 days.

It is further noteworthy that the results of the second experiment were highly consistent with the ones of the first experiment. Similar to the first experiment, the imagery group showed better performance than the control group in transfer, retention, and drawing scores, and moreover, the students' transfer and drawing scores were strongly related, with $r = .52$ in Experiment 1 and with $r = .59$ in Experiment 2. In line with these results, mediation analyses indicated that students' performance in the drawing test mediated the effect of imagery instruction on transfer performance in Experiments 1 and 2. These results confirm the idea that mental imagination fosters deep processing of the text content that contributes to the durability of these effects. Overall, the fact that the imagination effect can be demonstrated on both an immediate test and a delayed test and on measures of transfer, retention, and drawing points to its robustness. The strong effect sizes (many above $d = 0.80$) point to its practical importance.

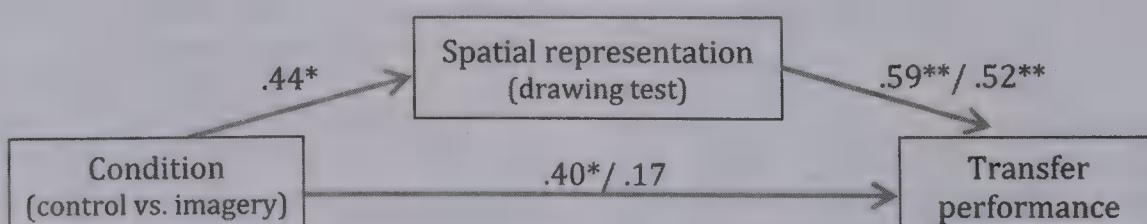


Figure 5. Mediation model in Experiment 2. * $p < .01$. ** $p < .001$. See the online article for the color version of this figure.

Theoretical Implications

The results are consistent with the idea that the act of imagining the spatial relations among elements described in an explanatory text can prime generative learning processes—including selecting relevant elements, organizing them into a coherent structure, and relating them to relevant prior knowledge. The act of imagining, when accomplished successfully, can help students build what Paivio (1986) called referential connections between corresponding words and images, similar to multimedia learning with presented drawings (Mayer, 2009). The positive effects of mental imagery on transfer performance suggest that classic theoretical concepts of mental imagery based on associative tasks (Paivio, 1986), mnemonic techniques (Atkinson, 1975), or recall of facts (Rasco et al., 1975) can be extended with respect to how mental imagery facilitates deep understanding of complex explanatory scientific text.

Theories that explain the multimedia principle (Butcher, 2014; Mayer, 2009) provide one starting point because constructing internal pictures in conjunction with corresponding text leads to similar effects on retention and transfer performance as processing external pictures in conjunction with corresponding text. This indicates that the benefits of external pictures are transferrable to internal pictures (i.e., pictures created through the learner's imagination). When learning from text and external pictures, students select, organize, and integrate words and corresponding pictures. When imagining text content, students also select and organize words and transform these words into mental images of the text content. This requires the students to draw referential connections between words and corresponding images. This process is an intrinsic component of the imagination process because the strategy cannot be applied without the students creating connections between the text and their mental images. This verbal-visual connection may be a main benefit of the imagination strategy that contributes to enhanced transfer performance on immediate and delayed tests. This process is based on the dual coding approach described in detail by Sadoski and Paivio (2013).

Theories of mental model building (Johnson-Laird, 1983) provide a second starting point because they specify the nature of the constructed representation. A mental model is a representation that preserves structural equivalence with the referential content. A mental model of the respiratory system therefore would represent the spatial relations among the components of the system and their interaction, for example, that the diaphragm is located beneath the lungs. When students imagine the structure and the functions of the respiratory components as described in the text, they are prompted to construct a representation that depicts the spatial relations of the system. This construction process is based on an interaction of information provided by the text and general knowledge such as existent models of the learner (Vandierendonck, Dierckx, & van der Beken, 2006).

Due to the spatial nature of the respiration process, students can animate and transform components of the system like a runnable mental model that should foster the students' ability to transfer their knowledge to new problems (Hegarty, 2004). One main characteristic of the imagery process is that students create their mental image step by step (Denis, 2008; Hegarty, 1992). This stepwise process helps the students to distinguish the separate components of the respiratory system and how they relate to one another. By contrast, when students perceive an

external picture, they initially perceive a holistic static image and need to put extra effort into processing the text and the picture in order to separate its particular components and functions. Consequently, the images constructed by the imagery group might have been more dynamic than the images constructed by the picture-and-imagery groups. Therefore, the imagery group may have found it easier to manipulate these images—an activity that is essential in transfer tasks.

Practical Implications

The results of the present two experiments point to important practical implications for using mental imagination to fostering deeper learning. In conjunction with the research reported in the introduction, these findings suggest that mental imagery is a powerful strategy to enhance transfer and retention performance. In particular, as a complement to the multimedia principle (Butcher, 2014; Mayer, 2009), we propose an imagination principle, which says that students learn better from explanatory scientific text when they are asked to imagine a coherent spatial representation depicting the relations among key elements in the text. An important boundary condition is that our students received very clear imagination prompts specifying what they should imagine to help them construct accurate images of the text content. Previous research has shown that students struggle to construct accurate mental images of a spatial outlay (Denis & Cocude, 1992). Thus, appropriate supports are critical to improving the quality of the mental images that will in turn affect test performance. A further advantage of mental imagery strategies is that they are easy to implement in reading education programs and educational settings. Contrary to related, better established strategies like drawing strategies, mental imagery strategies do not require the students to invest additional resources in externalizing their images (Leutner et al., 2009; Leutner & Schmeck, 2014).

Limitations and Future Directions

Some limitations of the study relate to materials, methodology, participants, and context. Concerning materials, our imagination treatment includes prompts to imagine the text along with the names of specific elements to include (e.g., thoracic cavity), so it is not possible to determine which aspects of the treatment—the imagining part, the elements part, or both—caused the improvement in learning. A potentially useful finding is that the imagination effect was mediated by the students' spatial representations. However, as mediation analysis does not allow us to draw causal inferences, further research is required to examine which aspects of the imagination treatment are most important in producing an improvement in transfer test performance. This methodological limitation could be addressed in future work by comparing instructions to "study" particular elements of the text versus to "imagine" particular elements of the text.

Furthermore, it should be noted that one prerequisite for the imagination strategy to function is that the text employed is written clearly enough so that students actually can imagine it because students have to rely solely on the text when constructing their mental pictures. Thus, it is worth investigating how to support students as they imagine and animate dynamic processes that are described in the text. A limitation of experiments that utilize

delayed tests, as we did in Experiment 2, is that it is difficult to collect reliable data on whether students studied the materials when they took the delayed test. Concerning participants, college students participated in the study so further research is required to determine whether the imagination effect can apply to other age groups and learners with different characteristics. Finally, concerning context, this was a short-term laboratory study, which produced promising results, so further work is needed to determine how an imagination strategy affects transfer problem solving in learning scientific material and how the imagination effect can be applied in courses involving multimedia learning.

References

Anderson, R. C., & Kulhavy, R. W. (1972). Imagery and prose learning. *Journal of Educational Psychology, 63*, 242–243. doi:10.1037/h0032638

Atkinson, R. C. (1975). Mnemotechnics in second-language learning. *American Psychologist, 30*, 821–828. doi:10.1037/h0077029

Baron, R. M., & Kenny, D. A. (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology, 51*, 1173–1182. doi:10.1037/0022-3514.51.6.1173

Best, R., Rowe, M., Ozuru, Y., & McNamara, D. S. (2005). Deep-level comprehension of science texts. *Topics in Language Disorders, 25*, 65–83. doi:10.1097/00011363-200501000-00007

Borst, G., & Kosslyn, S. M. (2012). Scanning visual mental images: Some structural implications, revisited. In V. Gyselinck & F. Pazzaglia (Eds.), *From mental imagery to spatial cognition and language* (pp. 19–42). London, England: Psychology Press.

Brigham, F. J., & Brigham, M. M. (1998). Using keyword mnemonics in general music classes: Cognitive psychology meets music history. *Journal of Research and Development in Education, 31*, 205–213.

Brünken, R., Seufert, T., & Paas, F. (2010). Measuring cognitive load. In J. L. Plass, R. Moreno, & R. Brünken (Eds.), *Cognitive load theory* (pp. 181–202). doi:10.1017/CBO9780511844744.011

Butcher, K. R. (2014). The multimedia principle. In R. E. Mayer (Ed.), *The Cambridge handbook of multimedia learning* (2nd ed.; pp. 174–205). New York, NY: Cambridge University Press.

Carney, R. N., & Levin, J. R. (1998). Coming to terms with the keyword method in introductory psychology: A “neuromnemonic” example. *Teaching of Psychology, 25*, 132–134. doi:10.1207/s15328023top2502_15

Cooper, G., Tindall-Ford, S., Chandler, P., & Sweller, J. (2001). Learning by imagining. *Journal of Experimental Psychology: Applied, 7*, 68–82. doi:10.1037/1076-898X.7.1.68

DeLeeuw, K. E., & Mayer, R. E. (2008). A comparison of three measures of cognitive load: Evidence for separable measures of intrinsic, extraneous, and germane load. *Journal of Educational Psychology, 100*, 223–234. doi:10.1037/0022-0663.100.1.223

Denis, M. (2008). Assessing the symbolic distance effect in mental images constructed from verbal descriptions: A study of individual differences in the mental comparison of distances. *Acta Psychologica, 127*, 197–210. doi:10.1016/j.actpsy.2007.05.006

Denis, M., & Cocode, M. (1989). Scanning visual images generated from verbal descriptions. *European Journal of Cognitive Psychology, 1*, 293–307. doi:10.1080/09541448908403090

Denis, M., & Cocode, M. (1992). Structural properties of visual images constructed from poorly or well-structured verbal descriptions. *Memory & Cognition, 20*, 497–506. doi:10.3758/BF03199582

Driskell, J. E., Copper, C., & Moran, A. (1994). Does mental practice enhance performance? *Journal of Applied Psychology, 79*, 481–492. doi:10.1037/0021-9010.79.4.481

Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving students’ learning with effective learning techniques: Promising directions from cognitive and educational psychology. *Psychological Science in the Public Interest, 14*, 4–58. doi:10.1177/1529100612453266

Etel, A., Scheiter, K., Schüller, A., Nyström, M., & Holmqvist, K. (2013). How a picture facilitates the process of learning from text: Evidence for scaffolding. *Learning and Instruction, 28*, 48–63. doi:10.1016/j.learninstruc.2013.05.002

Ekstrom, R. B., French, J. W., Harman, H. H., & Dermen, D. (1976). *Manual for kit of factor-referenced cognitive tests*. Princeton, NJ: Educational Testing Service.

Farah, M. J. (1984). The neurological basis of mental imagery: A componential analysis. *Cognition, 18*, 245–272. doi:10.1016/0010-0277(84)90026-X

Finke, R. A. (1985). Theories relating mental imagery to perception. *Psychological Bulletin, 98*, 236–259. doi:10.1037/0033-2909.98.2.236

Gambrell, L. B., & Jawitz, P. B. (1993). Mental imagery, text illustrations, and children’s story comprehension and recall. *Reading Research Quarterly, 28*, 264–276. doi:10.2307/747998

Ganis, G., Thompson, W. L., & Kosslyn, S. M. (2004). Brain areas underlying visual mental imagery and visual perception: An fMRI study. *Cognitive Brain Research, 20*, 226–241. doi:10.1016/j.cogbrainres.2004.02.012

Giesen, C., & Peeck, J. (1984). Effects of imagery instruction on reading and retaining a literary text. *Journal of Mental Imagery, 8*, 79–90.

Ginns, P., Chandler, P., & Sweller, J. (2003). When imagining information is effective. *Contemporary Educational Psychology, 28*, 229–251. doi:10.1016/S0361-476X(02)00016-4

Goolsby, R. D., & Sadoski, M. (2013). A theoretical approach to improving patient education through written material. *Annals of Behavioral Science and Medical Education, 19*, 14–18.

Graesser, A. C. (2007). An introduction to strategic reading comprehension. In D. McNamara (Ed.), *Reading comprehension strategies: Theories, interventions, and technologies* (pp. 3–26). Mahwah, NJ: Erlbaum.

Hegarty, M. (1992). Mental animation: Inferring motion from static displays of mechanical systems. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18*, 1084–1102. doi:10.1037/0278-7393.18.5.1084

Hegarty, M. (2004). Mechanical reasoning by mental simulation. *Trends in Cognitive Sciences, 8*, 280–285. doi:10.1016/j.tics.2004.04.001

Johansson, R., Holsanova, J., & Holmqvist, K. (2006). Pictures and spoken descriptions elicit similar eye movements during mental imagery, both in light and in complete darkness. *Cognitive Science, 30*, 1053–1079. doi:10.1207/s15516709cog0000_86

Johnson-Laird, P. N. (1983). *Mental models: Towards a cognitive science of language, inference and consciousness*. Cambridge, England: Cambridge University Press.

Jones, M. S., Levin, M. E., Levin, J. R., & Beitzel, B. D. (2000). Can vocabulary-learning strategies and pair-learning formats be profitably combined? *Journal of Educational Psychology, 92*, 256–262. doi:10.1037/0022-0663.92.2.256

Kester, L., Kirschner, P. A., & Van Merriënboer, J. J. G. (2005). The management of cognitive load during complex cognitive skill acquisition by means of computer-simulated problem solving. *British Journal of Educational Psychology, 75*, 71–85. doi:10.1348/000709904X19254

Kincaid, J. P., Fishburne, R. P., Rogers, R. L., & Chissom, B. S. (1975). *Derivation of new readability formulas (Automated Readability Index, Fog Count, and Flesch Reading Ease formula) for Navy enlisted personnel* (Research Branch Report 8–75). Memphis, TN: Chief of Naval Technical Training, Naval Air Station Memphis.

Klockars, A. J., & Sax, G. (1986). *Multiple comparisons*. Newbury Park, CA: Sage.

Kosslyn, S. M., Ball, T. M., & Reiser, B. J. (1978). Visual images preserve metric spatial information: Evidence from studies of image scanning.

Journal of Experimental Psychology: Human Perception and Performance, 4, 47–60. doi:10.1037/0096-1523.4.1.47

Kosslyn, S. M., Thompson, W. L., & Ganis, G. (2006). *The case for mental imagery*. doi:10.1093/acprof:oso/9780195179088.001.0001

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology*, 77, 1121–1134. doi:10.1037/0022-3514.77.6.1121

Kulhavy, R. W., & Swenson, I. (1975). Imagery instructions and the comprehension of text. *British Journal of Educational Psychology*, 45, 47–51. doi:10.1111/j.2044-8279.1975.tb02294.x

Leahy, W., & Sweller, J. (2005). Interactions among the imagination, expertise reversal, and element interactivity effects. *Journal of Experimental Psychology: Applied*, 11, 266–276. doi:10.1037/1076-898X.11.4.266

Leopold, C., Sumfleth, E., & Leutner, D. (2013). Learning with summaries: Effects of representation mode and type of learning activity on comprehension and transfer. *Learning and Instruction*, 27, 40–49. doi:10.1016/j.learninstruc.2013.02.003

Leutner, D., Leopold, C., & Sumfleth, E. (2009). Cognitive load and science text comprehension: Effects of drawing and mentally imagining text content. *Computers in Human Behavior*, 25, 284–289. doi:10.1016/j.chb.2008.12.010

Leutner, D., & Schmeck, A. (2014). The generative drawing principle in multimedia learning. In R. E. Mayer (Ed.), *The Cambridge handbook of multimedia learning* (2nd ed.; pp. 433–448). New York, NY: Cambridge University Press.

Levin, J. R., Morrison, C. R., McGivern, J. E., Mastropieri, M. A., & Scruggs, T. E. (1986). Mnemonic facilitation of text-embedded science facts. *American Educational Research Journal*, 23, 489–506. doi:10.3102/00028312023003489

MacKinnon, D. P., Lockwood, C. M., Hoffman, J. M., West, S. G., & Sheets, V. (2002). A comparison of methods to test mediation and other intervening variable effects. *Psychological Methods*, 7, 83–104. doi:10.1037/1082-989X.7.1.83

Mayer, R. E. (2009). *Multimedia learning* (2nd ed.). doi:10.1017/CBO9780511811678

Mayer, R. E., & Gallini, J. K. (1990). When is an illustration worth ten thousand words? *Journal of Educational Psychology*, 82, 715–726. doi:10.1037/0022-0663.82.4.715

Mayer, R. E., & Sims, V. K. (1994). For whom is a picture worth ten thousand words? Extensions of a dual coding theory of multimedia learning. *Journal of Educational Psychology*, 86, 389–401. doi:10.1037/0022-0663.86.3.389

Mayer, R. E., Steinhoff, K., Bower, G., & Mars, R. (1995). A generative theory of textbook design: Using annotated illustrations to foster meaningful learning of science text. *Educational Technology Research and Development*, 43, 31–41. doi:10.1007/BF02300480

McCormick, C. B., Levin, J. R., & Valkenaar, D. E. (1990). How do mnemonic and thematic strategies affect students' prose learning? *Reading Psychology*, 11, 15–31. doi:10.1080/0270271900110103

Paivio, A. (1965). Abstractness, imagery, and meaningfulness in paired-associate learning. *Journal of Verbal Learning and Verbal Behavior*, 4, 32–38. doi:10.1016/S0022-5371(65)80064-0

Paivio, A. (1969). Mental imagery in associative learning and memory. *Psychological Review*, 76, 241–263. doi:10.1037/h0027272

Paivio, A. (1975). Coding distinctions and repetition effects in memory. In G. H. Bower (Ed.), *The psychology of learning and motivation* (Vol. 9, pp. 179–214). doi:10.1016/S0079-7421(08)60271-6

Paivio, A. (1986). *Mental representations: A dual coding approach*. New York, NY: Oxford University Press.

Paivio, A. (2007). *Mind and its evolution: A dual coding theoretical approach*. Mahwah, NJ: Erlbaum.

Paivio, A., & Csapo, K. (1973). Picture superiority in free recall: Imagery or dual coding? *Cognitive Psychology*, 5, 176–206. doi:10.1016/0010-0285(73)90032-7

Pressley, G. M. (1976). Mental imagery helps eight-year-olds remember what they read. *Journal of Educational Psychology*, 68, 355–359. doi:10.1037/0022-0663.68.3.355

Rasco, R. W., Tennyson, R. D., & Boutwell, R. C. (1975). Imagery instructions and drawings in learning prose. *Journal of Educational Psychology*, 67, 188–192. doi:10.1037/h0077014

Raugh, M. R., & Atkinson, R. C. (1975). A mnemonic method for learning a second-language vocabulary. *Journal of Educational Psychology*, 67, 1–16. doi:10.1037/h0078665

Sadoski, M., Goetz, E. T., & Fritz, J. B. (1993). Impact of concreteness on comprehensibility, interest, and memory for text: Implications for dual coding theory and text design. *Journal of Educational Psychology*, 85, 291–304. doi:10.1037/0022-0663.85.2.291

Sadoski, M., Goetz, E. T., & Rodriguez, M. (2000). Engaging texts: Effects of concreteness on comprehensibility, interest, and recall in four text types. *Journal of Educational Psychology*, 92, 85–95. doi:10.1037/0022-0663.92.1.85

Sadoski, M., & Paivio, A. (2013). *Imagery and text: A dual coding theory of reading and writing* (2nd ed.). New York, NY: Taylor & Francis.

Sadoski, M., & Quast, Z. (1990). Reader response and long-term recall for journalistic text: The roles of imagery, affect, and importance. *Reading Research Quarterly*, 25, 256–272. doi:10.2307/747691

Schwamborn, A., Mayer, R. E., Thillmann, H., Leopold, C., & Leutner, D. (2010). Drawing as a generative activity and drawing as a prognostic activity. *Journal of Educational Psychology*, 102, 872–879. doi:10.1037/a0019640

Schworm, S., & Renkl, A. (2006). Computer-supported example-based learning: When instructional explanations reduce self-explanations. *Computers & Education*, 46, 426–445. doi:10.1016/j.compedu.2004.08.011

Shepard, R. N., & Cooper, L. A. (1982). *Mental images and their transformations*. Cambridge, MA: MIT Press.

Shepard, R. N., & Metzler, J. (1971, February 19). Mental rotation of three-dimensional objects. *Science*, 171, 701–703. doi:10.1126/science.171.3972.701

Sheskin, D. J. (2011). *Handbook of parametric and nonparametric statistical procedures* (5th ed.). Boca Raton, FL: Chapman & Hall/CRC.

Sobel, M. E. (1982). Asymptotic confidence intervals for indirect effects in structural equation models. In S. Leinhardt (Ed.), *Sociological methodology* (pp. 290–312). doi:10.2307/270723

Vandierendonck, A., Dierckx, V., & Van der Beken, H. (2006). Interaction of knowledge and working memory in reasoning about relations. In C. Held, G. Vosgerau, & M. Knauff (Eds.), *Mental models and the mind* (pp. 53–84). doi:10.1016/S0166-4115(06)80027-0

VanLehn, K., Graesser, A. C., Jackson, G. T., Jordan, P., Olney, A., & Rose, C. P. (2007). When are tutorial dialogues more effective than reading? *Cognitive Science*, 31, 3–62. doi:10.1080/03640210709336984

Van Merriënboer, J. J. G., & Ayres, P. (2005). Research on cognitive load theory and its design implications for e-learning. *Educational Technology Research and Development*, 53, 5–13.

Appendix A**Introduction and Paragraphs of the Text on the Respiratory System**

Topic	Text
Introduction	Respiration is the process that moves air in and out of the lungs. Through respiration oxygen is delivered to where it is needed in the body and carbon dioxide is removed from the body. Respiration involves three phases: inhaling, exchanging and exhaling. The respiratory process is controlled by the nervous system.
Structure of the Nervous System	The respiratory center is located in the rear, bottom part of the brain, near the back of the neck. The respiratory center of the brain is connected to a pathway of nerves that leads down from the spinal cord to connect with muscles controlling the diaphragm and rib cage.
Steps in the Nervous System to Control Breathing	When the brain detects the need for more oxygen in the bloodstream, the respiratory center in the brain sends out a signal to inhale. The signal moves along the pathway of nerves to muscles controlling the diaphragm and rib cage. When the brain detects the need for less carbon dioxide in the bloodstream, the respiratory center in the brain terminates the signal to inhale. The signal to inhale stops moving along the pathway of nerves to the muscles controlling the diaphragm and rib cage.
Structure of the Thoracic Cavity	The thoracic cavity is the space in the chest that contains the lungs. It is surrounded by the rib cage, which can move slightly inward or outward, and has the diaphragm on the bottom, which has a dome that can move downward. The main muscles involved in respiration are the diaphragm and the rib muscles. The diaphragm is located underneath the lungs. It lines the lower part of the thoracic cavity, sealing it off air-tight from the rest of the body. The rib muscles are attached to the ribs, which in turn encircle the lungs. When in the relaxed position, the ribs are slightly inward and the diaphragm dome curves upward.
Structure of the Airway System	From the nose and the mouth the windpipe leads to the bronchial tubes, which branch off into the right and the left lung. There they branch off into finer tubes.
Process of Inhaling	During inhaling, a signal from the brain to inhale causes the dome of the diaphragm to contract downward and the rib cage to move slightly outward creating more space in the thoracic cavity into which the lungs can expand. Air is drawn in through the nose or mouth, moves down through the windpipe and bronchial tubes to tiny air sacs in the lungs.
Structure of the Exchange System	Tiny grape-like air sacs, called alveoli, are grouped together in the lungs at the bronchial tubes. Each air sac is surrounded by tiny blood vessels called capillaries. On one side of the air sac the surrounding capillaries carry oxygen and on the other side they carry carbon dioxide. Oxygen-carrying capillaries connect air sacs to larger blood vessels called arteries and are represented as red because they contain an abundance of oxygen. Carbon-dioxide-carrying capillaries connect larger blood vessels called veins to the air sacs and are represented as blue because they contain an abundance of carbon dioxide.
Structure of the Circulatory System	Arteries (red blood vessels) run one-way from the lungs, through the heart, which is somewhat below the lungs, to the cells of the body. Arteries transport oxygen, which is used by the cells of the body to make energy. Veins (blue blood vessels) run one-way in the opposite direction from the cells of the body, through the heart, to the lungs. Veins transport carbon dioxide, which is a waste gas produced in the cells of the body. The heart is a pump that keeps the blood flowing in the veins and arteries.
Process of Exchanging	The exchange of oxygen and carbon dioxide takes place in the connection between air sacs and capillaries. Oxygen molecules in the inhaled air move to the capillaries running nearby, and carbon dioxide molecules move from the capillaries into the air sacs in the lungs. The capillaries carry the oxygen to arteries, which transport it, through the heart, to the cells of the body. At the same time, carbon dioxide travels in veins from the cells of the body, through the heart, to capillaries running next to the air sacs.
Process of Exhaling	The carbon-dioxide-rich air in the air sacs is drawn out of the lungs by exhaling. When the brain turns off the signal to inhale, the diaphragm and the rib muscles relax. The dome of the diaphragm moves upward again and the ribs move slightly inward. As a result, the thoracic cavity becomes smaller creating less room for the lungs. Air containing carbon dioxide is forced out of the lungs through the bronchial tubes and windpipe to the nose and mouth, where it leaves the body.

(Appendices continue)

Appendix B

Imagination Instructions for the Nine Text Paragraphs

Paragraph	Instruction
1	Please imagine the structure of the nervous system consisting of the brain, nerves, diaphragm, and rib muscles.
2	Please imagine the steps in the nervous system when the brain sends a signal to the diaphragm and rib muscles.
3	Please imagine the structure of the thoracic portion, consisting of the thoracic cavity, lungs, rib cage, and diaphragm.
4	Please imagine the structure of the airway portion, consisting of the nose, mouth, windpipe, bronchial tubes, lungs, and air sacs.
5	Please imagine the steps in the thoracic cavity and the airway when the diaphragm and rib muscles receive a signal to inhale.
6	Please imagine the structure of the exchange system consisting of air sacs, oxygen-carrying capillaries, carbon-dioxide-carrying capillaries, veins, and arteries.
7	Please imagine the structure of the circulatory system consisting of lungs, arteries, veins, heart, and cells of the body.
8	Please imagine the steps in the exchange system and the circulatory system for the process of exchanging.
9	Please imagine the steps in the thoracic cavity, airway, diaphragm and rib muscles for the process of exhaling.

Received July 22, 2013

Revision received May 1, 2014

Accepted May 3, 2014 ■

Matching Learning Style to Instructional Method: Effects on Comprehension

Beth A. Rogowsky
Bloomsburg University of Pennsylvania

Barbara M. Calhoun
Vanderbilt University

Paula Tallal
Rutgers University and University of California,
San Diego

While it is hypothesized that providing instruction based on individuals' preferred learning styles improves learning (i.e., reading for visual learners and listening for auditory learners, also referred to as the *meshing hypothesis*), after a critical review of the literature Pashler, McDaniel, Rohrer, and Bjork (2008) concluded that this hypothesis lacks empirical evidence and subsequently described the experimental design needed to evaluate the meshing hypothesis. Following the design of Pashler et al., we empirically investigated the effect of learning style preference with college-educated adults, specifically as applied to (a) verbal comprehension aptitude (listening or reading) and (b) learning based on mode of instruction (digital audiobook or e-text). First, participants' auditory and visual learning style preferences were established based on a standardized adult learning style inventory. Participants were then given a verbal comprehension aptitude test in both oral and written forms. Results failed to show a statistically significant relationship between learning style preference (auditory, visual word) and learning aptitude (listening comprehension, reading comprehension). Second, participants were randomly assigned to 1 of 2 groups that received the same instructional material from a nonfiction book, but each in a different instructional mode (digital audiobook, e-text), and then completed a written comprehension test immediately and after 2 weeks. Results demonstrated no statistically significant relationship between learning style preference (auditory, visual word) and instructional method (audiobook, e-text) for either immediate or delayed comprehension tests. Taken together, the results of our investigation failed to statistically support the meshing hypothesis either for verbal comprehension aptitude or learning based on mode of instruction (digital audiobook, e-text).

Keywords: learning styles, listening and reading comprehension, audiobooks, e-text

Teaching to individuals' perceived learning styles in hopes that they will achieve greater academic success is common practice within the field of education. Not only does the learning styles concept have widespread acceptance among educators (Dekker, Lee, Howard-Jones, & Jolles, 2012) but also it is accepted among the general public (Pashler, McDaniel, Rohrer, & Bjork, 2008).

This article was published Online First July 28, 2014.

Beth A. Rogowsky, College of Education, Bloomsburg University of Pennsylvania; Barbara M. Calhoun, Vanderbilt Brain Institute, Vanderbilt University; Paula Tallal, Center for Molecular and Behavioral Neuroscience, Rutgers University, and Center for Human Development, University of California, San Diego.

We are grateful to Audible, Inc., who provided the digital audiobook and e-text materials used in this study. We are also grateful to Susan Rundle who provided the Building Excellence Learning Style Inventory gratis. Additionally, we would like to thank the Temporal Dynamics of Learning Center, a National Science Foundation (NSF) Science of Learning Center funded by NSF Grant SBE-0542013.

Correspondence concerning this article should be addressed to Beth A. Rogowsky, College of Education, 2213 McCormick Center, Bloomsburg University of Pennsylvania, Bloomsburg, PA 17815. E-mail: brogowsk@bloomu.edu or brogowsky@gmail.com

The learning style literature, as well as learning style inventories, differs widely in the way that learning styles are conceived and assessed (see Coffield, Moseley, Hall, & Ecclestone, 2004, and Pashler et al., 2008, for review). For example, in the Gregorc Style Delineator (Gregorc, 1982), learning styles are defined by perception (concrete or abstract) and ordering (sequential or random). The Kolb's Learning Style Inventory (1985) emphasizes experiential learning and includes accommodating, diverging, converging, and assimilating styles. Herrmann's Brain Dominance Instrument (1996) categorizes learners as theorists (cerebral, left: the rational self), organizers (limbic, left: the safe-keeping self), innovators (cerebral, right: the experimental self), and humanitarians (limbic, right: the feeling self). Dunn and Dunn's Learning Styles Inventory (Dunn, Dunn, & Price, 1989) concentrates on modality-specific strengths and weaknesses (e.g., visual, auditory, tactile, and kinesthetic processing). In the current study, we focused on verbal comprehension, specifically, the extent to which verbal comprehension may be influenced by the modality of input: auditory (digital audio) or visual (e-text).

While the learning styles literature has been extensively discussed and reviewed, there are considerably more theoretical and descriptive discussions on this topic than there are empirical stud-

ies. For example, Cassidy (2004) described the central themes and issues surrounding learning styles and the many instruments available for the measurement of learning styles with the goal of promoting research in the field. Kozhevnikov (2007) presented a literature review on cognitive styles, which served as a basis for the author's theory that suggests that cognitive styles represent heuristics that can be identified at multiple levels of information processing, from perceptual to metacognitive, and that individuals can be grouped according to the type of regulatory function they exert. Sternberg, Grigorenko, and Zhang (2008) divided learning and thinking into two basic styles: ability based and personality based, and advocated that both are important for instruction and assessment. They argued that teachers need to take into consideration differences in how students learn and think and design instruction accordingly to obtain optimal instructional outcomes.

The importance of evaluating students' learning styles and developing instructional methods that teach to specific learning styles has gained considerable support in the field of education, with many organizations and companies offering professional development courses for teachers and educators focused on the topic of learning styles. For this reason, Pashler, McDaniel, Rohrer, and Bjork (2008) were charged with reviewing the empirical evidence pertaining to the importance of assessing and teaching to students' learning styles for the journal *Psychological Science in the Public Interest*. In their review, they define *learning styles* as "the concept that individuals differ in regard to what mode of instruction or study is most effective for them. . . . The most common—but not the only—hypothesis about the instructional relevance of learning styles is the meshing hypothesis, according to which instruction is best provided in a format that matches the preferences of the learner (e.g., for a 'visual learner,' emphasizing visual presentation of information; p. 105)." After reviewing the literature, they found that while there is evidence that, if asked, both children and adults indicate preferences as to how they favor information be presented to them, and there is also evidence that people have specific aptitudes for processing different types of instruction, there is limited empirical evidence as to whether providing instruction in an individual's preferred learning style (i.e., listening for those with an auditory learning style or reading for those with a visual learning style) improves learning. Furthermore, they also concluded that the definitive study showing that individuals with a preferred auditory learning style learn better when listening rather than reading, and conversely, that those with a preferred visual learning style learn better when reading rather than listening, had not been conducted.

Given the lack of credible validation of learning-styles-based instruction, Pashler et al. (2008) described a three-step experimental design of the study that would need to be conducted, as well as the pattern of data that would need to be found, in order to conclude empirically that learning is significantly improved when individuals receive instruction tailored to their asserted learning style. In Step 1, participants must be divided into groups on the basis of their learning style. In Step 2, participants from each group must be randomly assigned to receive one of multiple instructional methods. In Step 3, participants must complete an assessment of the material that is the same for all students. For the learning styles meshing hypothesis to be supported, data analysis must reveal a specific type of interaction between learning style and instructional

method. That is, learning is optimal when individuals receive instruction in their preferred learning style, and the instructional method that proves most effective for individuals with one learning style is not the most effective method for individuals with a different learning style.

Pashler et al. (2008) also pointed out that educators as well as the general public fail to distinguish between learning style preferences and learning aptitude. They stated that "[t]here is, after all, a commonsense reason why the two concepts could be conflated: Namely, different modes of instruction might be optimal for different people because different modes of presentation exploit the specific perceptual and cognitive strengths of different individuals, as suggested by the meshing hypothesis" (pp. 109–110). However, the relationship between learning style preference and learning aptitude, specifically as it relates to the meshing hypothesis and verbal comprehension, has not been established empirically.

In 2012, Dekker, Lee, Howard-Jones, and Jolles reported that 94% of educators believed that students perform better when they receive information in their preferred learning style (e.g., auditory, visual, kinesthetic). Given this continued widespread belief and the influence of learning styles on educational practice, coupled with the importance of verbal comprehension on educational outcomes, we conducted an investigation of the meshing hypothesis as it pertains to verbal aptitude and learning. We implemented the methodology and analyses proposed by Pashler et al. (2008) in order to directly test the following two research questions:

1. What is the extent to which learning style preferences (auditory, visual) equate to learning aptitudes (listening comprehension, reading comprehension)?

2. What is the extent to which learning style preferences and/or learning aptitudes predict how much an individual comprehends and retains based on mode of instruction (audiobook, e-text)?

In the first research question, we investigated the relationship between learning style preferences (as measured by a standardized learning style inventory) and learning aptitudes (as measured by a listening and reading comprehension assessment). Specifically, as applied to the relationship between verbal aptitude and learning style preference, the meshing hypothesis predicts that (a) there will be a positive correlation between auditory learning style preference and listening comprehension, (b) there will be a positive correlation between visual word learning style preference and reading comprehension, and (c) individuals with a visual learning style preference will comprehend better when they read rather than listen, and conversely, individuals with an auditory learning style preference will comprehend better when they listen rather than read.

In the second research question, we investigated the extent to which learning style preferences (auditory, visual) and/or learning aptitudes (listening comprehension, reading comprehension) predict how much an individual will learn and retain based on two different modes of instruction (audiobook, e-text). Specifically, the meshing hypothesis predicts that individuals with a visual learning style preference learn more when they read e-text rather than when they listen to an audiobook, and conversely, individuals with an auditory learning style preference learn more when they listen to an audiobook rather than read e-text. Analogous predictions would be expected with regards to the relationship between listening

comprehension aptitude and learning from an audiobook and reading comprehension aptitude and learning from e-text.

Method

Participants

In order to be included in this study, participants had to meet the following inclusionary criteria: age 25–40 years; college educated (bachelor's degree only); native speakers of English; normal hearing and vision (with correction); and no self-reported history of neurological or learning impairments. Potential participants outside this age range, who had more advanced degrees beyond a bachelor's degree, who had not graduated from college, or who had a history of neurological or learning disabilities were excluded. Based on these criteria, 121 participants from the New York City metropolitan area were selected. Of the total population of 121 subjects, 62 were male and 59 were female. The mean age of the participants was 30.6 years ($SD = 4.4$). All participants completed 16 years of education. This study examined the two research questions. For Research Question 1, the entire population of 121 individuals participated. These 121 individuals were then randomly assigned to four groups. Two of these groups (61 participants) participated in Research Question 2. The remaining participants who had been randomly assigned to the other two groups participated in a different study that was not focused on learning styles. The 61 participants in Research Question 2 were randomly assigned to a listening condition ($n = 30$) or a reading condition ($n = 31$). The analyses of Research Question 2 focused only on those participants who could be categorically classified as having an auditory or visual word learning style and who were randomly assigned to either a listening or reading condition. The final four subgroups included in Research Question 2 analyses were listening condition with auditory learning style ($n = 11$), listening condition with visual word learning style ($n = 10$), reading condition with auditory learning style ($n = 10$), and reading condition with visual word learning style ($n = 10$).

This study was conducted in accordance with the prescribed standards of the institutional review board of Rutgers University–Newark. All participants provided informed consent and were financially compensated for their participation.

Learning Styles Assessment

Prior to on-site testing, participants completed an online standardized learning styles preference inventory. Pashler et al. (2008) identified the Dunn and Dunn learning styles model as being one of the most popular learning styles assessment tools because of the constructs included as well as the broad age range of the assessments offered—from children as young as 3 years old through adults. For this study, we selected the adult version, the Building Excellence (BE) Online Learning Styles Assessment Inventory for ages 17 and older (Rundle & Dunn, 2010). The BE Learning Styles Inventory is a self-administered online survey that requires 20–25 min for completion. The assessment measure asks participants to decide if they *strongly disagree, disagree, are uncertain, agree, or strongly agree* after reading statements indicating, for example, whether the respondent remembers new information better by reading about it or by listening to a discussion about it (Rundle &

Dunn, 2010). The BE Learning Styles Inventory assesses individual learning and productivity styles based upon six domains: perceptual, psychological, environmental, physiological, emotional, and sociological. The perceptual domain is subdivided into the following six elements: auditory (input), visual picture, visual word, tactile, kinesthetic, and auditory verbal (output). The BE Learning Styles Inventory provides an individual's strengths and weaknesses pertaining to these six possible perceptual learning styles. For each learning style preference, individuals are placed into one of five bins that are continuous, ranging from very weak to very strong. For example, the five bins for auditory are classified as (1) *strong less auditory*, (2) *moderate less auditory*, (3) *it depends*, (4) *moderate more auditory*, and (5) *strong more auditory*. Within each bin, there is a 3-point range with the exception of Bin 3 (*it depends*) that has a 5-point range, for a total of 17 possible placements along the continuum for each perceptual element. For the purpose of this study, we focused only on those elements (auditory and visual word) that most relate to listening and reading comprehension, respectively.

The BE Learning Styles Inventory provides personalized reports that convert an individual's numerical score into instructional recommendations. For example, if a participant scores *strong less auditory* or *moderate less auditory* (Bin 1 or Bin 2, respectively/ corresponding Placements 1–6), the recommendation prescribed by the BE Learning Styles Inventory would be that because listening is not a strength, the participant should rely on a stronger style when learning new material. If a participant scores *strong less visual word* or *moderate less visual word* (Bin 1 or Bin 2, respectively/ corresponding Placements 1–6), the recommendation prescribed would be that because reading is not a strength, the participant should rely on a stronger style when learning new and difficult information. If a participant scores *it depends* in either auditory or visual word (Bin 3/ corresponding Placements 7–11), the BE Learning Styles recommendation acknowledges that the participant is indifferent to the modality. He or she is encouraged to use one of his or her strengths when learning new information. If a participant scores *moderate more auditory/visual* or *strong more auditory/visual* (Bin 4 or Bin 5, respectively/ corresponding Placements 12–17), the individual is advised to use that style most of the time when learning. While the automated computer scoring system generated scores and reports for each participant, the participants were not informed about the purpose of the study or given access to their scores or reports or given any feedback from this survey.

In this study, learning styles data were analyzed using two different scoring procedures. For correlation and regression analyses, data were analyzed using the full standard continuous 17-point scoring method provided by the BE Learning Styles Inventory. Three variables were used: BE auditory (range = from +1 to +17), BE visual word (range = from +1 to +17), and the difference between BE auditory and BE visual word (range from -17 to +17). In this study, participants' BE auditory scores ranged from +2 to +17, their visual word scores ranged from +5 to +17 and the difference between BE auditory minus BE visual word scores ranged from -11 to +8. ($M = -0.92$, $SD = 3.98$).

In addition, in order to follow the analysis prescribed by Pashler et al. (2008), which addressed the meshing hypothesis directly, individuals must first be divided into groups on the basis of their preferred learning style. For this purpose, participants were clas-

sified *categorically* as having primarily either an auditory or visual word learning style. We used the five bin categories provided by the BE Learning Styles Inventory: *strong less auditory/visual word* = 1; *moderate less auditory/visual word* = 2; *it depends* = 3; *moderate more auditory/visual word* = 4; and *strong more auditory/visual word* = 5. According to the BE Learning Styles Inventory, only participants who scored *moderate* to *strong more auditory* (either a 4 or 5) as well as *it depends* or *moderate* to *strong less visual word* (3, 2, 1) were instructed to use the auditory modality "much of the time." For the purposes of this analysis, these participants were classified as having an auditory learning style ($n = 37$). Similarly, only participants who scored *moderate* to *strong more visual word* (either a 4 or 5) as well as *it depends* or *moderate* to *strong less auditory* (3, 2, 1) were instructed to use the visual word modality "much of the time." These participants were classified as having a visual word learning style ($n = 31$). Of the 121 individuals who participated in Research Question 1, 53 participants could not be categorically classified as either auditory or visual word learners and, as such, were not included in the analyses that required categorical classification.

Verbal Comprehension Aptitude Measure

The goal of this study was to determine the extent to which learning style preference (auditory, visual) and/or verbal aptitude (listening comprehension, reading comprehension) relates to the effectiveness of instructional method (audiobook, e-text). Because there is no standardized assessment designed to directly compare listening and reading comprehension aptitude in adults, we developed a verbal comprehension aptitude test in both a listening and reading format using matched passages from two equivalent forms of the fourth edition of the Gray Oral Reading Test (GORT-4). GORT-4 is a standardized assessment measure composed of leveled passages that objectively measure oral reading rate, accuracy, fluency, and comprehension, as well as alerts to possible learning exceptionalities (Weiderholt & Bryant, 2000). GORT-4 (age range: from 6.0 to 18.11 years) consists of 13 passages that become increasingly difficult as the examinee progresses. After pilot testing all passages in college-educated adults, we selected passages 9, 10, 11, and 13 for use in this study. Passages 1–8 and 12 did not provide sufficient individual differences in our college-educated population and were not included. None of the individuals who participated in pilot testing were included in the current study. The selected passages ranged from 148 to 167 words ($M = 158$, $Mdn = 157$). Each passage was followed by five comprehension questions. To assess listening comprehension, we converted the selected passages from Form B of the GORT-4 into a digital audio recording. A professional audiobook narrator, who read at a steady pace and with natural intonation, recorded the passages. We will refer to this assessment as the *Listening Aptitude Test (L-AT)*. To assess reading comprehension, we asked each participant to read the selected passages from Form A of the GORT-4 silently. This assessment will be referred to as the *Reading Aptitude Test (R-AT)*.

Each of the 121 participants in this study was tested on both the L-AT and the R-AT. The order in which the L-AT and the R-AT were taken was counterbalanced to reduce the chance that the order of testing would adversely influence the results. Half of the participants completed the R-AT and then L-AT, where they read

the first four passages and then listened to the remaining four passages; the other half of the participants completed the L-AT and then R-AT, where they listened to the first four passages and then read the remaining four passages. Participants read each passage silently from a computer screen or listened through headphones to a digital audio recording.

Immediately after they read or listened to each passage, participants answered the five corresponding multiple-choice questions for that passage. Note that part of the answer from one of the questions on the R-AT was accidentally omitted. Therefore, data could only be collected from 19 of the 20 questions. To assure that the R-AT and the L-AT remained equivalent, the comparable question from the L-AT was also deleted from all analyses. The protocol designed by Pashler et al. (2008) to assess the meshing hypothesis requires individuals to complete an assessment that is the same for all participants. All participants answered the comprehension questions in the same (written) format. We chose to focus on this response format because most tests of comprehension are administered in writing. The program required a response for each question before the participant could proceed to the next question. Participants were not permitted to re-read or re-listen to any passage nor were they allowed to use the passage as a reference when answering the questions. No feedback was given.

Instructional Unit

Two modes of instruction were investigated for the same unit (audiobook, e-text). The content used across both of these instructional conditions was the preface and Chapter 17 of the nonfiction book, *Unbroken: A World War II Story of Survival, Resilience, and Redemption*, written by Laura Hillenbrand and read by Edward Hermann. The total content contained 3,184 words. Forty-eight multiple-choice questions were designed to assess the participants' comprehension. These 48 questions will be referred to as the *Unbroken* comprehension test.

The question set was developed by a certified teacher of English (B.R.), who serves on the Pennsylvania State Standardized Assessment Panel where she reviews reading assessment items for content, rigor, alignment, bias, and universal and technical design. Questions were piloted for difficulty on a sample of 10 individuals meeting the eligibility requirements for participants in this study but were not participants in the study. The *Unbroken* comprehension test was given twice, once immediately following completion of the passage (Time 1) and again 2 weeks later (Time 2).

Procedure

After completing the Listening Aptitude Test (L-AT) and Reading Aptitude Test (R-AT), participants were randomly assigned to one of two instructional conditions for the *Unbroken* portion of the study. Participants in each of the instructional conditions received the preface and Chapter 17 of *Unbroken*, presented in one of two different formats. In the audiobook condition, participants used headphones to listen to both the preface and Chapter 17 of *Unbroken* presented on an electronic tablet in digital audiobook format. In the e-text condition, participants read both the preface and Chapter 17 of *Unbroken* presented on an electronic tablet in e-text format. A research assistant pre-cued the e-text or audio, as well as monitored the participant to assure that there were no

interruptions and that the participant understood how to use the equipment, was on-task, and did not extend reading/listening beyond the prescribed passages. Prior to administration of the passage for the audio condition, the volume was adjusted to a comfortable level. The audio condition lasted 16 min 24 s and was read at a pace of 149 words per minute. Participants in the e-text condition read at their own pace without time restraint. The replaying/fast-forwarding of audio and the re-reading/skipping of text were prohibited. The research assistant monitored participants' compliance.

Upon completion of Chapter 17, participants proceeded immediately (Time 1) to take the *Unbroken* comprehension test and answer 48 questions derived from the preface and Chapter 17. Participants were not allowed to use the e-text or digital audiobook as a reference. Each question was individually displayed in written text only on a computer screen, as is common in standard testing practices. The online multiple-choice assessment required a response for each question before the examinee could proceed to the next question. No feedback was given. In addition to the online, on-site immediate comprehension assessment (Time 1), participants completed the same multiple-choice assessment online 2 weeks later (Time 2) in order to evaluate their retention of the information in the story.

Results

Analyses for Research Question 1

Research Question 1 addresses the extent to which learning style preferences (auditory, visual word) as measured by the BE Learning Style Inventory equate to learning aptitudes (listening comprehension, reading comprehension) as measured by the L-AT and the R-AT.

To evaluate the equivalence of the L-AT and the R-AT for assessing comprehension aptitude in this population ($N = 121$), we calculated a paired-samples t test comparing the mean of the L-AT ($M = 13.9$, $SD = 3.4$) to the mean of the R-AT ($M = 12.8$, $SD = 2.8$). A significant difference was found, $t(120) = 3.54$; $p < .01$. The mean of the L-AT was significantly higher than the mean on the R-AT with an effect size of Cohen's $d = 0.32$. Although this difference was not ideal, it is important to note that the main hypothesis pertaining to the interaction between learning style and mode of instruction does not require that the R-AT and L-AT measures be equivalent.

Analyses using categorical learning style variables to predict learning aptitude: Implementing the Pashler et al. (2008) method. Pashler et al. (2008) prescribed a specific methodology for assessing the meshing hypothesis that requires that participants be categorically classified into two discrete learning styles (auditory learners or visual word learners). To follow this methodology explicitly, participants were classified into two discrete learning style categories: auditory learners ($n = 37$) or visual word learners ($n = 31$) as described in the Methods section.

A one-way multivariate analysis of variance (MANOVA) was calculated examining the effects of learning style preference groups (auditory, visual word) on the L-AT and R-AT scores to determine if learning style preference (auditory, visual word) predicts listening or reading comprehension aptitude. A significant effect of aptitude test (L-AT vs. R-AT) was found, $F(1, 66) =$

12.7; $p < .05$, with an effect size $\eta^2 = 0.16$, indicating that participants performed significantly better on one aptitude test (L-AT: $M = 14.1$, $SD = 3.5$) than on the other aptitude test (R-AT: $M = 12.8$, $SD = 2.9$). There was also a significant effect of learning styles preference (auditory vs. visual word), $F(1, 66) = 6.9$; $p < .05$, with an effect size $\eta^2 = 0.09$, indicating that participants in one learning styles preference group (visual word: $M = 14.4$, $SD = 4.0$) performed significantly better than the participants in the other learning styles preference group (auditory, $M = 12.6$, $SD = 3.6$). There was not a significant aptitude test (L-AT, R-AT) by learning styles preference (auditory, visual word) interaction, $F(1, 66) = 0.34$; $p > .05$. Further inspection of these results using one-way analyses of variance (ANOVAs) show that overall, participants in the visual word learning style group scored significantly higher on the L-AT, $F(1, 66) = 5.48$, $p < .05$ ($M = 15.16$, $SD = 3.10$), than participants in the auditory learning style group ($M = 13.22$; $SD = 3.65$). Participants in the visual word learning style group also scored significantly higher on the R-AT, $F(1, 66) = 4.91$; $p < .05$ ($M = 13.58$, $SD = 2.50$), than participants in the auditory learning style group ($M = 12.08$; $SD = 2.99$). These results indicate that participants who had a visual word learning style preference were significantly better at *both* listening and reading comprehension, compared to those who had an auditory learning style preference.

According to Pashler et al. (2008), acceptable evidence in support of the meshing hypothesis would show a crossover between two learning style preference groups (auditory, visual word) and listening and reading comprehension aptitude (L-AT, R-AT), as shown in Figure 1A. Figure 1B shows an example taken from Pashler et al. (2008) of one form of unacceptable evidence for the meshing hypothesis, where both auditory and visual word learning style preference groups score higher on the same method, and hence there is no crossover. Figure 1C shows the data from the current study. As shown in Figure 1C, contrary to the crossover pattern that would be expected to support the meshing hypothesis, the auditory and visual word learning style preference groups *both* scored higher on listening comprehension than on reading comprehension. It is important to note that not only was the L-AT performance better for both groups but also the superiority of the L-AT over the R-AT was similar for both groups. According to Pashler et al. (2008), this pattern corresponds to one example of unacceptable evidence in support of the meshing hypothesis.

However, classification of participants into two discrete groups reduces the sensitivity of continuous variables and also reduced the sample size by including only those participants who had a clear auditory or visual word learning style preference. To mitigate these concerns, we performed a final series of correlation and step-wise multiple regression analyses ($n = 121$) to evaluate whether there was a significant relationship between learning style preference (BE Learning Styles Inventory) and learning aptitude (L-AT, R-AT). For these analyses, variables from the BE Learning Styles Inventory based on the continuous 17-point standard BE scoring system were used to predict verbal comprehension aptitude scores on the L-AT and R-AT.

Correlation and regression analyses for Research Question 1. The relationship between learning style preference (BE Learning Styles Inventory) and listening and reading comprehension aptitude (L-AT and R-AT) was evaluated by a series of simple correlation analyses as well as stepwise multiple regression anal-

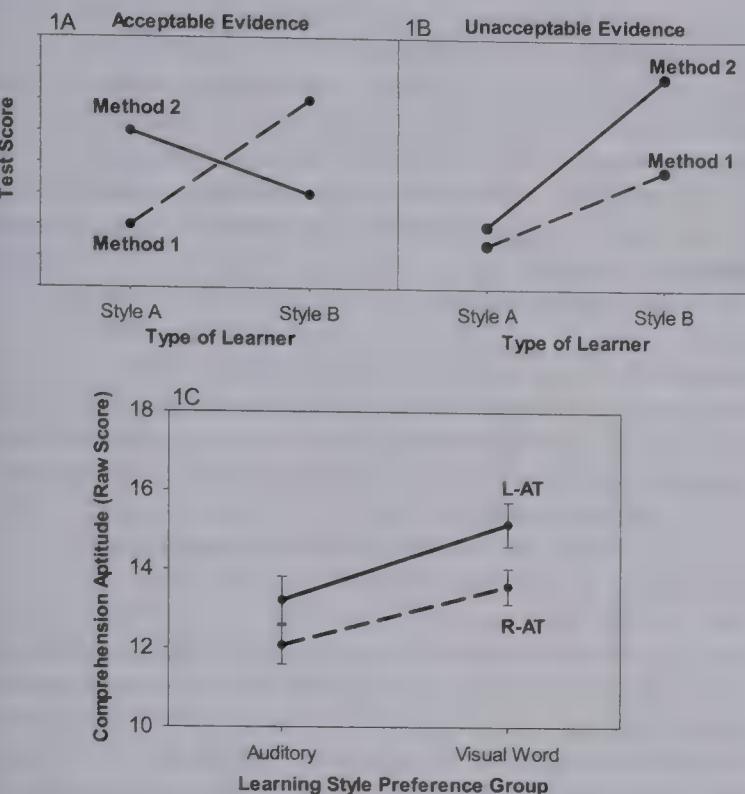


Figure 1. Graph A displays the pattern of evidence required to support the meshing hypothesis while Graph B displays one of several patterns of evidence that would constitute unacceptable evidence (according to Pashler et al., 2008). Graph C displays the results from the current study, which show that there is no crossover effect. Bars represent standard errors. The 95% confidence interval (CI) for the Listening Aptitude Test (L-AT) ranged from 12.10 to 14.33 for participants with an auditory learning preference and from 13.94 to 16.39 for participants with a visual learning preference. The 95% CI for the Reading Aptitude Test (R-AT) ranged from 11.17 to 12.99 for participants with an auditory learning preference and from 12.58 to 14.58 for participants with a visual learning preference.

yses. For these analyses, the following variables from the BE Learning Styles Inventory were used based on the 17-point standard BE scoring system: BE auditory score, BE visual word score, and the difference between the BE auditory score and the BE visual word score (BE auditory score – BE visual word score), to predict verbal comprehension aptitude. Verbal comprehension aptitude outcomes of interest included (a) predicting listening aptitude (L-AT raw score), (b) predicting reading aptitude (R-AT raw

score), and (c) predicting the difference between listening and reading aptitude (L-AT – R-AT raw score). Table 1 presents the means and standard deviations for each of the BE learning styles and verbal comprehension variables as well as the correlation matrix for these variables, and Table 2 presents the results of the multiple regression analyses.

Predicting listening comprehension aptitude from learning style preference scores. The meshing hypothesis predicts a positive correlation between learning style preference and aptitude. That is, if auditory learning style equates to listening comprehension aptitude, as auditory learning style preference scores increase, listening comprehension aptitude score would also increase. As seen in Table 1, contrary to expectation based on the meshing hypothesis, the correlation between auditory learning style preference (based on the BE auditory score) and listening comprehension (based on the L-AT score) was negative ($-.31, p < .01$). To further test whether other learning style variables influence listening comprehension aptitude, we calculated multiple linear regression analyses to determine the extent to which participants' listening comprehension aptitude (L-AT) could be predicted based on their BE auditory learning style score, BE visual word learning style score, and the difference between their BE auditory and BE visual word scores. As seen in Table 2, a significant regression equation was found, $F(1, 119) = 12.96, p < .001$, with an R^2 of .10. The only BE learning style variable that contributed significantly to the listening comprehension score was the BE auditory learning style score. This single variable contributed a correlation coefficient of $R = .31, R^2 = .10 (SE = 3.28), p < .001$. Participants' predicted listening comprehension score is equal to 17.20 (constant) $+ -0.30$ (BE auditory learning style score), indicating a negative relationship between the BE auditory learning style score and the listening comprehension aptitude score. The coefficient model shows that for every 1 point that the BE auditory learning style score decreased, the listening comprehension score increased 0.30 points. BE visual word learning style score and the difference between BE auditory learning style score and BE visual word learning style score failed to contribute any significant variance beyond that already accounted for by the BE auditory learning style score. This analysis demonstrated that only the BE auditory learning style score accounted for a significant portion of the listening comprehension variance. However, contrary to what would be predicted by the meshing hypothesis, this relationship

Table 1
Descriptive Statistics and Correlation Matrix for the Predictor Variables Entered into the Multiple Regression Aptitude Model for Research Question 1

Variable	<i>M</i>	<i>SD</i>	1	2	3	4	5	6
1. Listening aptitude	13.87	3.44	—	.46	.66	-.31**	-.14	-.21*
2. Reading aptitude	12.81	2.78		—	-.37	-.24**	-.04	-.19*
3. Difference between listening and reading aptitude	1.67	3.28			—	-.13	-.11	-.054
4. BE auditory learning style	10.98	3.55				—	.081	.85**
5. BE visual word learning style	11.89	2.13					—	-.46**
6. Difference between BE auditory and visual word learning styles	-0.92	3.99						—

Note. $n = 121$; BE = Building Excellence.

* $p < .05$. ** $p < .01$.

Table 2

Coefficients for the Significant Predictor Variables Entered into the Multiple Regression Model for Listening Aptitude, Reading Aptitude, and Difference Between Listening and Reading Aptitude for Research Question 1

Variable	B	SEB	β	R	R ²
Listening aptitude				.31	.10
Constant	17.20	0.97	**		
BE auditory learning style score	-0.30	0.084	-.31**		
Reading aptitude				.24	.06
Constant	14.84	0.81	**		
BE auditory learning style score	-0.18	0.070	-.24**		
Difference between listening and reading aptitude	—	—	—		

Note. The following predictor variables were entered into the model for (a) listening aptitude, (b) reading aptitudes, and (c) difference between listening and reading aptitudes; Building Excellence (BE) auditory learning style score; BE visual word learning style score; and the difference between BE auditory and BE visual word score. Only the variables listed above made significant contributions to these models. No variables made significant contributions to the model for difference between listening and reading aptitude. Shown are the coefficients (B), the standard error of the coefficients (SEB), as well as standardized coefficient (β), and the correlation. $N = 121$.

* $p < .05$. ** $p < .01$.

was negative. That is, as auditory learning style preference increased, performance on a listening aptitude test decreased.

Predicting reading comprehension aptitude from learning preference scores. The meshing hypothesis would predict that if visual word learning style preference equates to reading comprehension aptitude, as participants' visual word learning style preference score increased, their reading comprehension aptitude score would also increase. As shown in Table 1, the correlation between visual word learning style preference (based on the BE visual word score) and reading comprehension (based on the R-AT score) was neither positive nor significant (-.04). To further test whether other learning style variables influence reading comprehension aptitude, we calculated a multiple linear regression analysis to determine the extent to which participants' reading comprehension aptitude (R-AT) could be predicted based on their BE auditory learning style score, BE visual word learning style score, and the difference between their BE auditory and BE visual word scores. As seen in Table 2, a significant regression equation was found, $F(1, 119) = 7.01$, $p < .01$, with an R^2 of .06. However, the only BE variable that contributed significantly to the reading comprehension score was the BE auditory learning style score. This single variable contributed a correlation coefficient of $R = .24$, $R^2 = .06$ ($SE = 2.72$), $p < .01$. Participants' predicted reading comprehension score is equal to 14.84 (constant) + -0.18 (BE auditory learning style score), indicating a negative relationship between BE auditory learning style score and reading comprehension aptitude score. The coefficient model shows that for every 1 point that the BE auditory learning style score decreased, the reading comprehension score increased 0.18 points. Contrary to the assumption that a visual verbal learning style preference would predict higher reading scores, neither the BE visual word learning style score nor the difference between the BE auditory learning style score and BE

visual word learning style score contributed any significant variance beyond that already accounted for by the BE auditory learning style score. This analysis demonstrated that auditory learning style was also the only significant predictor of reading comprehension scores, and this relationship was again negative.

Predicting the difference between listening comprehension aptitude and reading comprehension aptitude from learning preference scores. The meshing hypothesis would predict that individuals who have a stronger auditory learning style preference would also have a higher listening versus reading comprehension aptitude score, and conversely, those who have a stronger visual word learning style preference would also have a higher reading versus listening comprehension aptitude score. A multiple linear regression was calculated to determine the extent to which participants' *difference between listening comprehension aptitude (L-AT)* and *reading comprehension aptitude (R-AT)* could be predicted based on their BE auditory learning style score, BE visual word learning style score, and the difference between their BE auditory and BE visual word scores. This regression analysis most completely tests the meshing hypothesis, which not only predicts a simple relationship between learning style preference and comprehension aptitude but also and more specifically predicts that individuals with different learning styles will perform differentially with different modes of input. The results of this analysis failed to support this prediction. None of the variables (BE auditory learning style score, BE visual word learning style score, or the difference between BE auditory and BE visual word scores) contributed significantly to the difference between listening comprehension aptitude and reading comprehension aptitude.

Discussion of analyses for Research Question 1. Pashler et al. (2008) pointed out that learning style preferences and learning aptitudes are often considered to be overlapping constructs. After all, it seems intuitive that individuals who prefer to listen would perform better on a test of listening than reading comprehension and, conversely, those who prefer reading would perform better on a test of reading than listening comprehension. This relationship is referred to as the *meshing hypothesis*. Research Question 1 was designed as an empirical test of this hypothesis, as it pertains to verbal comprehension aptitude. Participants completed the BE Learning Styles Inventory as well as both a listening and a reading comprehension aptitude test. A series of analyses were calculated to determine the extent to which auditory and visual word learning style variables predicted listening and/or reading comprehension aptitude. Both a continuous score (based on the 17-point scale established by the BE Learning Style Inventory) and a categorical classification of participants as either an auditory learner or visual word learner were included in these analyses. This categorical classification was based on the 5-point BE scale and included only those participants with a strong difference between their auditory and visual word reading preference scores. Regardless of whether continuous or categorical scores from the BE Learning Styles Inventory were used, the results were consistent in failing to provide statistically significant support for the meshing hypothesis. Contrary to the expectations predicted by the meshing hypothesis, that a high visual word learning style score would be the best predictor of a high reading comprehension aptitude score, and conversely, that a high auditory learning style score would be the best predictor of a high listening comprehension aptitude score, auditory learning style proved to be the only significant predictor

of *both* reading and listening comprehension scores, and in both cases this relationship was negative. That is, as individuals' auditory learning style preference scores increased, their performance on *both* the listening and reading comprehension aptitude tests decreased. Thus, the results using the BE learning style preference scores both as a continuous scale as well as a discrete categorical measure of auditory and visual word learning style preference fail to demonstrate a significant positive relationship between (a) auditory learning style preference and listening aptitude, (b) visual word learning style preference and reading aptitude, or (c) a differential effect of learning style preference on performance on a listening compared with a reading comprehension aptitude test. These findings fail to support the construct that an individual's learning style (auditory, visual word) is positively correlated with their listening and reading aptitude. Taken together, these data fail to provide statistical support for the meshing hypothesis, at least as it pertains to verbal comprehension (listening vs. reading) aptitudes.

Limitations of analyses for Research Question 1. One potential concern for the analyses reported for Research Question 1 was that the L-AT and R-AT were not matched for difficulty. It is important to emphasize that the tests developed for this study to assess listening and reading comprehension aptitude, while derived from two equivalent forms of a standardized, normed reading test (GORT-4), were not given in the standard format on which these norms were based. The main question of interest is whether there was a *differential* pattern of results for auditory compared with visual word learners when listening compared with reading. In this study, both the auditory and the visual word learning style preference groups scored higher on the listening than on the reading comprehension aptitude test. This could be an indication that the listening version of the test was easier than the written version. While equivalent scores would have been more ideal, it is important to note that it is the *pattern* of the results, rather than the absolute values, that is critical in addressing the meshing hypothesis. As shown in Figure 1C, the difference in listening compared with reading performance resulted in parallel lines for the auditory learners compared with the visual word learners. That is, while participants classified by the BE Learning Style Inventory as auditory learners did, indeed, score higher on a comprehension test when they listened to versus read passages, participants classified as visual word learners showed a similar pattern; that is, they also scored higher on this same comprehension test when they listened to versus when they read the test passages—and to the same degree. Taken in context, the results from Research Question 1 are contrary to the pattern that would be expected based on the meshing hypothesis, at least as it applies to tests of listening and reading comprehension aptitude. However, this research question does not address the issue of whether learning and retention of "real-world," nonfiction material, presented using different instructional methods, is affected by an individual's preferred learning style. This was the focus of Research Question 2.

Analyses for Research Question 2

Research Question 2 addresses the extent to which learning style preferences (as measured by the BE Learning Style Inventory) and/or learning aptitudes (as measured by the L-AT and R-AT) predict how much an individual comprehends and retains based on

mode of instruction (audiobook, e-text) as measured by the *Unbroken* comprehension test.

The validity of the *Unbroken* comprehension test was evaluated to assure that results obtained from this test were an accurate measure of comprehension. To do this, the comprehension score of each participant on the *Unbroken* comprehension test at Time 1 was compared with the same participant's total comprehension aptitude score (total verbal comprehension aptitude = L-AT + R-AT). A Pearson correlation coefficient was calculated. A positive correlation was found, $r(119) = 0.59$, $p < .001$, indicating that there was a significant relationship between participants' scores on the total verbal comprehension aptitude test and participants' scores on the *Unbroken* comprehension test at Time 1. This analysis showed that participants who had higher comprehension scores as indicated by the total verbal comprehension aptitude test also had higher comprehension scores on the *Unbroken* test at Time 1, providing construct validity for this test. Next, a Pearson correlation coefficient was calculated for the relationship between the *Unbroken* comprehension test at Time 1 and Time 2. A strong positive correlation was found, $r(118) = 0.86$, $p < .01$, indicating a significant linear relationship between the two variables. Participants who performed well on the *Unbroken* comprehension test at Time 1 performed well on this same test at Time 2. This linear relationship indicates strong test-retest reliability for the *Unbroken* comprehension test.

A 2 (modes of instruction) \times 2 (time) mixed-design ANOVA was calculated to evaluate the effects of mode of instruction (audiobook, e-text) and time (Time 1, Time 2) on the *Unbroken* comprehension test scores. There was a main effect of time (Time 1 vs. Time 2), $F(1, 58) = 37.3$; $p < .05$. However, there was no significant main effect for mode of instruction, $F(1, 58) = 0.25$; $p > .05$. In addition, there was not a significant mode of instruction by time interaction, $F(1, 58) = 0.08$; $p > .05$. These results indicate that there was no difference in difficulty on the *Unbroken* comprehension test when presented by audiobook versus e-text. Moreover, all participants performed better in both instructional conditions at Time 1 than Time 2.

Analyses using categorical learning style variables to predict learning via audiobook versus e-text mode of instruction at Time 1: The Pashler et al. (2008) method. When the meshing hypothesis is applied to education theory and practice, it is assumed that learning will be more effective when material is presented in an instructional mode that meshes with the individual's preferred learning style. Pashler et al.'s (2008) meshing hypothesis pertaining to learning style preference and modes of instruction predicts that individuals with a visual learning style preference will comprehend better when they read rather than listen, and conversely, individuals with an auditory learning style preference will comprehend better when they listen rather than read. The Pashler et al. (2008) roadmap for evaluating the meshing hypothesis empirically begins by dividing participants into distinct learning style preference groups. Therefore, for this analysis, participants were classified as auditory or visual word learners based on their BE Learning Styles Inventory results, as described in the Methods section.

Results from Research Question 1 showed that there were significant differences in reading and listening aptitude for study participants based on their learning style preference. Recall that participants with a visual word learning style preference achieved

both higher listening and reading aptitude scores than participants in the auditory learning style preference group. As a result, to control for any effect of potential differences in total verbal comprehension aptitude, based on the random assignment to instructional condition in Research Question 2, we conducted all analyses both with and without co-varying out the effect of total reading and listening aptitude. No significant differences were found with or without the covariance. As such, only the ANOVA results are reported. Table 3 shows the *Unbroken* comprehension test raw scores (total number correct out of 48) by learning style preference group (auditory, visual word) and instructional condition (audiobook, e-text) at Time 1 and Time 2. A between-subjects 2 (learning style preference) \times 2 (mode of instruction) ANOVA was performed using these data to examine the effect of different learning style preferences (auditory, visual word) and different modes of instruction (audiobook, e-text) on the *Unbroken* comprehension test scores at Time 1. The results of this analysis showed that the main effect for learning style preference was significant, $F(1, 37) = 6.11; p < .05$, indicating a significant difference between participants with an auditory learning style preference ($M = 30.57; SD = 5.89$), and those with a visual word learning style ($M = 34.40; SD = 3.33$). This demonstrates that *Unbroken* comprehension at Time 1 was affected by learning style preference, with the participants with visual word learning styles performing better. However, the main effect for instructional condition (audiobook, e-text) was not significant, $F(1, 37) = .15; p > .05$, with no significant difference in performance on the *Unbroken* comprehension test at Time 1 between participants in the audiobook condition ($M = 32.10; SD = 6.00$) and those in the e-text condition ($M = 32.80; SD = 4.16$). Finally, the interaction between instructional condition (audiobook, e-text) and learning style preference (auditory, visual word) was not significant, $F(1, 37) = 0.42; p > .05$, indicating that providing instruction in a mode that matched an individual's learning style preference did *not* result in significantly better learning. Figure 2C shows the results of this analysis.

According to Pashler et al. (2008), acceptable evidence in support of the meshing hypothesis would show a crossover between the two learning style preference groups and two modes of instruction, as shown in Figure 2A. Figure 2B shows an example from Pashler et al. (2008) of unacceptable evidence for the meshing hypothesis, where both auditory and visual word learning style preference groups score higher on the same method of instruction, and hence there is no crossover. As seen in Figure 2C, contrary to

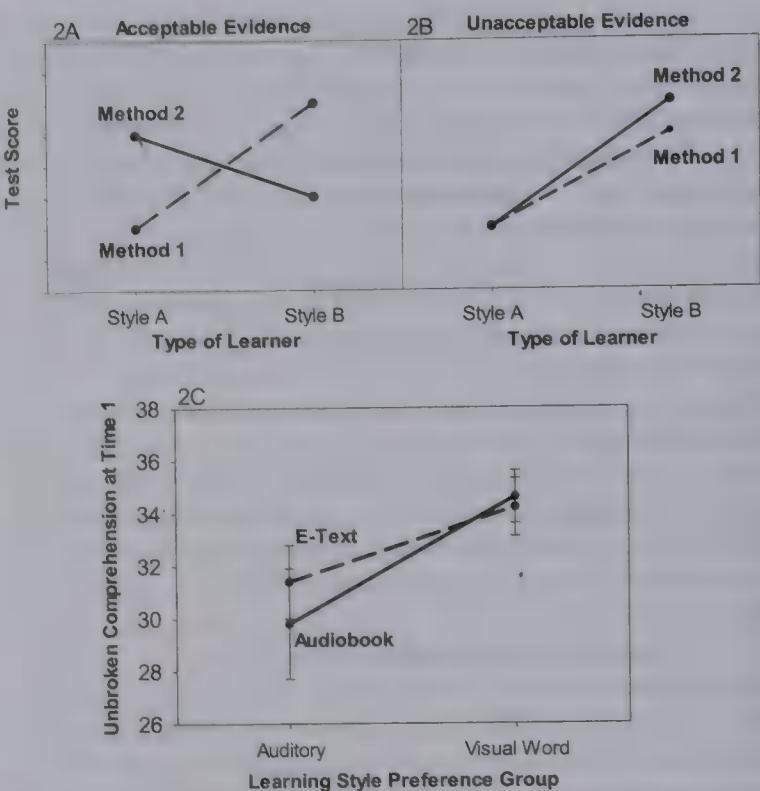


Figure 2. Examples of (A) acceptable evidence and (B) unacceptable evidence for the meshing hypothesis (according to Pashler et al., 2008). Graph C displays the results from this study and corresponds to one of Pashler et al.'s (2008) examples of unacceptable evidence. Error bars represent standard errors. The 95% confidence interval (CI) for the audiobook condition ranged from 26.82 to 32.81 for participants with an auditory learning preference and from 31.46 to 37.74 for participants with a visual learning preference. The 95% CI for the e-text group ranged from 28.26 to 34.54 for participants with an auditory learning preference and from 31.06 to 37.34 for participants with a visual learning preference.

the crossover pattern that would be expected to support the meshing hypothesis, results from the current study show there is minimal difference based on instructional condition for participants in either the auditory or visual word learning style preference groups. According to Pashler et al. (2008), this pattern corresponds with one example of unacceptable evidence in support of the meshing hypothesis.

Two-week retention (Time 2). It is possible that presenting instruction in a mode that meshes with an individual's learning style may affect longer term retention of information. To address this possibility, we calculated a 2 (learning style preference) \times 2 (mode of instruction) ANOVA to examine the long-term (2-week) effect of instructional condition (audiobook, e-text) and learning style preference (auditory, visual word) on *Unbroken* comprehension test scores at Time 2. The results at Time 2 parallel those found at Time 1. That is, the main effect for learning style preference was significant, $F(1, 37) = 9.18; p < .05$, indicating that *Unbroken* comprehension test scores at Time 2 were affected by learning style preference. Participants with an auditory learning style preference performed significantly more poorly ($M = 27.33; SD = 5.74$) than those with a visual word learning style preference ($M = 32.25; SD = 4.23$). The main effect of group was not significant, $F(1, 37) = .03; p > .05$, with no significant difference between participants using an audiobook ($M = 29.52; SD = 6.55$), and those using e-text ($M = 29.95; SD = 4.51$). Finally, the

Table 3
Unbroken Comprehension Test Raw Scores (Total Number Correct Out of 48) by Learning Style Preference Group and Instructional Condition at Time 1 and Time 2

Instructional condition	Building Excellence learning preference					
	Auditory			Visual word		
	<i>n</i>	<i>M</i>	<i>SD</i>	<i>n</i>	<i>M</i>	<i>SD</i>
Audiobook (Time 1)	11	29.82	7.10	10	34.60	3.27
E-text (Time 1)	10	31.40	4.43	10	34.20	3.55
Audiobook (Time 2)	11	26.64	7.06	10	32.70	4.30
E-text (Time 2)	10	28.10	4.07	10	31.80	4.34

interaction between instructional condition and learning style preference was not significant, $F(1, 37) = 0.54$; $p > .05$, indicating that providing instruction in a mode that matched an individual's learning style preference did not result in significantly better retention.

There were no significant interactions between learning style preference and mode of instruction based on a categorical classification of participants into two discrete groups, those with an auditory and those with a visual word learning style. However, classification of participants into two discrete groups reduced the sensitivity of continuous variables and also reduced the sample size by including only those participants who had a clear auditory or visual word learning style preference. To mitigate these concerns, we conducted a final series of correlation and stepwise multiple regression analyses to evaluate whether there was a significant relationship among learning style preference (BE Learning Styles Inventory), learning aptitude (L-AT, R-AT), and mode of instruction (digital audiobook, e-text). For these analyses, variables from the BE Learning Styles Inventory based on the continuous 17-point standard BE scoring system and verbal aptitude scores based on the L-AT and R-AT were used to predict (a) learning outcomes from the audiobook mode of instruction and (b) learning outcomes from the e-text mode of instruction (Table 4).

Correlation and regression analyses for Research Question 2.

Predicting audiobook learning outcomes from learning style preference and verbal aptitude scores at Time 1. The meshing hypothesis predicts a positive correlation between learning style preference and instructional mode. That is, as auditory learning style preference scores increase, learning outcomes via the audiobook mode of instruction, but not via the e-text mode of instruction, would also increase. As seen in Table 4, counter to what would be predicted by the meshing hypothesis, when the Pearson correlation was calculated examining the relationship between auditory learning style preference and learning from an audiobook, a weak negative correlation that was not significant was found, $r(28) = -.30$, $p > .05$. When the Pearson correlation was calculated examining the relationship between visual word learning

style preference and learning from an audiobook, the results were similar; a weak negative correlation that was not significant was found, $r(28) = -.24$, $p > .05$.

Similarly, to further test whether any other learning style or aptitude variables influenced learning outcomes from the audiobook mode, a multiple linear regression was calculated to determine the extent to which participants' learning of nonfiction material presented in audiobook format (instructional condition) could be predicted based on their learning style preference (BE auditory, BE visual word, and the difference between BE auditory and BE visual word scores) as well as comprehension aptitude (R-AT, L-AT, and verbal comprehension aptitude difference). As seen in Table 5, a significant regression equation was found, $F(1, 28) = 16.18$, $p < .001$. Listening aptitude (L-AT) was the only variable that contributed significantly to the comprehension of the material presented via the audiobook condition. This single variable contributed a correlation coefficient of 0.61, $R^2 = .37$ ($SE = 5.43$), $p < .001$. The regression equation showed that comprehension of material in the audiobook instructional condition was equal to 15.71 (constant) + 1.15 (listening comprehension aptitude), indicating a positive relationship between listening aptitude, and auditory instruction. The coefficient model shows that for every 1 point the audiobook comprehension increased, the listening aptitude score increased 1.15 points. This analysis demonstrated that only a component of aptitude (L-AT) contributed significantly to the variance in learning based on auditory instruction. BE auditory learning style score, BE visual word learning style score, and the difference between BE auditory and BE visual word score, as well as reading aptitude (R-AT), and the difference between listening aptitude and reading aptitude (L-AT-R-AT), failed to contribute any significant variance beyond that already accounted for by the listening aptitude score.

Predicting e-text learning outcomes from learning style preference and verbal aptitude scores at Time 1. The meshing hypothesis predicts a positive correlation between learning style preference and instructional mode. That is, as visual word learning style preference scores increase, learning outcomes via the e-text

Table 4
Descriptive Statistics and Correlation Matrix for the Predictor Variables Entered into the Multiple Regression Model for the Audiobook and e-Text Instructional Conditions at Time 1 as Described in Research Question 2

Condition	<i>n</i>	<i>M</i>	<i>SD</i>	Unbroken comprehension test results at Time 1					
				1	2	3	4	5	6
Audiobook									
1. Listening aptitude	30	30.87	6.70	.61**	.45*	.19	-.30	-.24	-.14
2. Reading aptitude	30	13.17	3.52		.46*	.58**	-.40	-.08	-.30
3. Difference between listening and reading aptitude	30	12.30	3.25			-.46**	-.42*	-.03	-.34
4. BE auditory learning style	30	0.87	3.54				-.01	-.05	.02
5. BE visual learning style	30	10.50	3.95					-.04	.88**
6. Difference between BE auditory and visual learning styles	30	11.80	2.23						-.30**
E-text									
1. Listening aptitude	31	31.35	5.40	.70**	.45*	.39*	-.25	.05	-.24
2. Reading aptitude	31	14.42	3.74		.43*	.72**	-.40*	-.01	-.33
3. Difference between listening and reading aptitude	31	13.10	2.72			-.31	-.20	.19	-.26
4. BE auditory learning style	31	1.32	3.54				-.27	-.16	-.14
5. BE visual learning style	31	10.97	3.70					-.05	.86**
6. Difference between BE auditory and visual learning styles	31	11.97	2.30						-.56**

Note. BE = Building Excellence.

* $p < .05$. ** $p < .01$.

Table 5
Coefficients for the Significant Predictor Variables Entered into the Multiple Regression Model for Audiobook and E-Text Learning at Time 1 as Described in Research Question 2

Variable	B	SEB	β	R	R^2	N
Audiobook						
Learning						
Constant	15.71	3.89	**	.61	.37	30
E-text						
Learning						
Constant	16.80	2.86	**	.70	.49	31
Listening aptitude	1.15	0.29	.61**			
Listening aptitude	1.01	0.19	.70**			

Note. The following predictor variables were entered into the model for audiobook and e-text learning: listening aptitude; reading aptitude; difference between listening and reading aptitude; Building Excellence (BE) auditory learning style score; BE visual word learning style score; and the difference between BE auditory and BE visual word score. Only the variables listed above made significant contributions to the model. Shown are the coefficients (B), the standard error of the coefficients (SEB), as well as standardized coefficient (β), and the correlation. $N = 30$ (audiobook); $N = 31$ (e-text).

* $p < .05$. ** $p < .01$.

mode of instruction, but not via the audiobook mode of instruction, would also increase. As shown in Table 4, when the Pearson correlation coefficients were calculated, there was a weak positive correlation between visual word learning style preference and learning from e-text, $r(29) = .05$, $p > .05$, and a weak negative correlation between auditory word learning style preference and learning from e-text ($r(29) = -.24$, $p > .05$). However, neither correlation was significant.

To further test whether any learning style or aptitude variables influence learning outcomes from the e-text mode of instruction, a multiple linear regression was calculated to determine the extent to which participants' learning of nonfiction material presented in e-text format (instructional condition) could be predicted based on their learning style preference (BE auditory, BE visual word, and the difference between BE auditory and BE visual word scores) as well as comprehension aptitude (R-AT, L-AT, and verbal comprehension aptitude difference). As shown in Table 5, a significant regression equation was found, $F(1, 29) = 27.67$, $p < .001$ with an R^2 of .49. Contrary to what would be predicted by the meshing hypothesis, however, the only variable that contributed significantly to the learning of the material presented in the e-text condition was listening comprehension aptitude (L-AT). This single variable contributed a correlation coefficient of $R = .70$, $R^2 = .49$ ($SE = 3.93$), $p < .001$. The regression equation showed that comprehension of material in the e-text instructional condition was equal to 16.80 (constant) + 1.01 (L-AT), indicating a positive relationship between listening comprehension aptitude and learning from e-text instruction. The coefficient model shows that for every 1 point e-text learning increased, listening comprehension aptitude increased by 1.01 points. This analysis demonstrated that only listening comprehension aptitude (L-AT) contributed significantly to the variance in learning material presented via e-text instruction. BE auditory learning style score, BE visual word learning style score, and the difference between BE auditory and BE visual word score, as well as R-AT, total verbal comprehen-

sion aptitude, and the verbal comprehension aptitude difference failed to contribute any significant variance in learning beyond that already accounted for by the L-AT.

Predicting audiobook and e-text learning outcomes from learning style preference scores only at Time 1. When both verbal comprehension aptitude and learning style preference variables were entered into multiple regression analyses to predict learning via either audiobook or e-text modes of instruction, only aptitude measures proved to significantly predict learning outcomes. In a final attempt to find a significant relationship between learning style preference and effects of instructional mode on learning, we conducted a regression analysis using *only* learning style preference variables (BE auditory, BE visual word, and the difference between BE auditory and BE visual word scores) to predict (a) audiobook and (b) e-text learning outcomes. The results of these analyses failed to provide any statistically significant support for the meshing hypothesis in that none of the BE learning style preference variables accounted for a statistically significant amount of variance for either audiobook ($p > .05$) or e-text ($p > .05$) learning outcomes.

Predicting audiobook and e-text learning outcomes from learning style preference and verbal aptitude scores at Time 2. Even if learning style preferences do not affect immediate learning of material based on mode of instruction (audiobook, e-text), it is possible that presenting instruction in a mode that meshes with an individual's learning style may affect longer term retention of information. Just as was done using the categorical variables for learning style preference, all analyses were repeated using the continuous variables based on the 2-week retention data obtained at Time 2. The results of these analyses are shown in Tables 6 and 7. As can be seen by directly comparing the correlation matrices obtained at Time 1 (Table 4) with those obtained at Time 2 (Table 6), as well as the multiple regression models obtained at Time 1 (Table 5) with those obtained at Time 2 (Table 7), the results were very similar at Time 2 to those found at Time 1. The only significant correlation found between audiobook learning and auditory learning style preference was found at Time 2, and this correlation was negative ($-.39$, $p < .05$). Similarly, results from the stepwise multiple regression analyses were similar at Time 2 to those found at Time 1; only aptitude scores positively predicted audiobook and e-text learning, with no significant learning preference variables entering the model (Table 7). Thus, similar to the results pertaining to immediate learning obtained at Time 1, the results obtained at Time 2 failed to provide any statistically significant evidence that showed that providing individuals with instruction in a mode that meshes with their learning style preference results in significantly better long-term retention of information.

Discussion of analyses for Research Question 2. Research Question 2 investigated the meshing hypothesis as it pertains to mode of instruction. Specifically, the meshing hypothesis predicts that participants with an auditory learning style preference will learn material better when instruction is presented via a listening mode than when it is presented via a written mode and, conversely, those with a visual word learning style preference will learn material better after having read it rather than having listened to it. An ANOVA was calculated to determine if the experiment provided any statistically significant evidence that showed that the

Table 6

Descriptive Statistics and Correlation Matrix for the Predictor Variables Entered into the Multiple Regression Model for Audiobook and E-Text Learning at Time 2 as Described in Research Question 2

Condition	n	M	SD	Unbroken comprehension test results at Time 2					
				1	2	3	4	5	6
Audiobook learning	30	28.20	7.76	.70**	.58**	.16	-.39*	-.25	-.21
1. Listening aptitude	30	13.17	3.52		.46*	.58**	-.40*	-.08	-.30
2. Reading aptitude	30	12.30	3.25			-.46**	-.42*	-.03	-.34
3. Difference between listening and reading aptitude	30	0.87	3.54				-.01	-.05	.02
4. BE auditory learning style	30	10.50	3.95					-.04	.88**
5. BE visual learning style	30	11.80	2.23						-.52**
6. Difference between BE auditory and visual learning styles	30	-1.30	4.62						
E-text learning	30	29.13	5.85	.66**	.52**	.32	-.23	.19	-.28
1. Listening aptitude	30	14.60	3.66		.54*	.70**	-.36*	.06	-.32
2. Reading aptitude	30	12.97	2.67			-.21	-.29	.13	-.30
3. Difference between listening and reading aptitude	30	1.63	3.15				-.17	-.05	-.12
4. BE auditory learning style	30	10.80	3.64					-.12	.87**
5. BE visual learning style	30	11.87	2.27						-.60**
6. Difference between BE auditory and visual learning styles	30	-1.07	4.53						

Note. BE = Building Excellence.

*p < .05. **p < .01.

method most effective for instructing individuals with one learning style is *not* the most effective method for individuals with a different learning style. The results of these analyses failed to provide empirical support for the meshing hypothesis. No significant interactions were found between learning style preference (auditory, visual word) and instructional method (digital audiobook, e-text) for either immediate learning or 2-week retention of verbal information.

A second series of simple and multiple regression analyses were conducted using continuous variables of both learning style preference as well as verbal comprehension aptitude. When both learning style and verbal comprehension aptitude variables were pitted against each other in multiple regressions to predict learning

via either digital audiobook or e-text, only the aptitude variables accounted for a significant amount of variance in learning. When only learning style variables were entered into these multiple regression analyses, they failed to account for a significant amount of variance in learning. Thus, regardless of scoring method used (categorical or continuous), the results from Research Question 2 failed to find a significant interaction between learning style preferences (auditory, visual word) and instructional method (digital audiobook, e-text) on learning or retention of information from a nonfiction text.

General Discussion

According to Pashler et al.'s (2008) recent review of the learning styles literature, there is widespread belief among educators and the general public alike that individuals learn better when they are presented instruction in the modality that capitalizes on their learning style preference. Pashler et al. (2008) focused on the extent to which auditory and visual learning style preferences influence verbal comprehension. Specifically, they focused on the meshing hypothesis that proposes that individuals with a visual learning style preference will learn more when information is presented to them in a written format, and conversely, those with an auditory learning style preference will learn more when instruction is presented to them in a listening format. They also pointed out that the meshing hypothesis may have led to the belief that learning style preferences and learning aptitudes for verbal comprehension are similar constructs. Their review of the literature led them to conclude, however, that there is little empirical evidence to support a direct relationship between learning style preferences (auditory, visual) and either (a) verbal comprehension aptitude (listening vs. reading) or (b) differential learning outcomes based on different modes of instruction (e.g., audiobook vs. e-text). However, they also concluded that the definitive study had not been conducted, and therefore, they prescribed a detailed roadmap for the experimental methodology that would be needed to address these important issues empirically as well as explicit examples of

Table 7

Coefficients for the Significant Predictor Variables Entered into the Multiple Regression Model for Learning From Audiobook and E-Text at Time 2, as Described in Research Question 2

Variable	B	SEB	β	R	R^2	n
Audiobook						
Comprehension				.76	.57	30
Constant	2.65	4.41				
Listening aptitude	1.21	0.31	.55*			
Reading aptitude	0.78	0.34	.33**			
E-text						
Comprehension				.66	.43	40
Constant	13.77	3.42	**			
Listening aptitude	1.05	0.23	.66**			

Note. The following predictor variables were entered into the model for audiobook and e-text learning: listening aptitude; reading aptitude; difference between listening and reading aptitude; Building Excellence (BE) auditory learning style score; BE visual word learning style score; and the difference between BE auditory and BE visual word score. Only the variables listed above made significant contributions to the model. Shown are the coefficients (B), the standard error of the coefficients (SEB), as well as standardized coefficient (β), and the correlation. N = 30 (audiobook); N = 30 (e-text).

*p < .05. **p < .01.

the patterns of data that would either support or refute the meshing hypothesis.

We conducted an investigation of the meshing hypothesis with college-educated adults following the research methods laid out by Pashler et al. (2008) to address two research questions. In Research Question 1, we used these methods to assess the extent to which an individual's learning style preference (auditory, visual word) was consistent with his or her learning aptitude for verbal comprehension (listening, reading). In Research Question 2, we used these methods to assess the extent to which an individual's learning style preference (auditory, visual-word) differentially affected how much they would learn and retain from nonfiction text presented using two different modes of instruction (digital audiobook, e-text).

Results from Research Question 1 showed that differences in preferred learning style (auditory, visual word) were *not* found to significantly predict differences in learning aptitude (listening vs. reading comprehension). That is, there were no statistically significant results that showed that individuals with stronger auditory learning style preferences had higher listening comprehension aptitude than reading aptitude or, conversely, that individuals with stronger visual word learning style preferences had higher reading than listening aptitude. Instead, participants classified with a preferred visual word learning style outperformed those classified as having a preferred auditory learning style on both the listening and reading comprehension aptitude tests. These results show that learning style preference and aptitude are not comparable constructs. Thus, the results from Research Question 1 failed to provide statistically significant support for the meshing hypothesis, at least as it pertains to the relationship between learning style preference (auditory, visual word) and verbal comprehension aptitude (listening, reading), respectively.

Similar to the results from Research Question 1, the results from Research Question 2 also failed to provide statistically significant empirical evidence supporting the meshing hypothesis, either for immediate learning or long-term retention of information presented via two different modes of instruction (audiobook, e-text). Regardless of whether categorical or continuous measures of learning styles were used or which method of analysis (ANOVA, simple correlations, multiple regression analyses) was chosen, there were no significant findings that showed that providing instruction to individuals in a mode that meshed with their preferred learning style resulted in better learning or retention of information compared with instructing them in their nonpreferred mode.

In conclusion, at least for verbal comprehension, no statistically significant evidence was found in this investigation to support the construct (a) that learning style is equivalent to learning aptitude or (b) that providing instruction in the modality that meshes with an individual's preferred learning style will result in significantly better learning or retention than presenting the same instruction in an individual's nonpreferred learning style.

Overall Limitations of the Study

One potential limitation in interpreting the results of Research Question 1 was that the L-AT and R-AT proved not to be matched for difficulty. Both the auditory and the visual word learning style preference groups scored higher on the listening than on the

reading comprehension aptitude test. This could be an indication that the listening version of the test was easier than the written version. We pointed out that while equivalent scores on the L-AT and the R-AT would have been more ideal, it is the *pattern* of the results, rather than the absolute values, that is critical in addressing the meshing hypothesis. That is, the main question of interest is whether there is a *differential* pattern of results for participants with an auditory compared with a visual word learning style preference in respect to listening compared with reading aptitude, and the analyses showed that there was not. The results from Research Question 2 also addressed this issue. Recall that in this case there was *no* significant main effect of condition (audiobook, e-text) on performance on the *Unbroken* comprehension test. There was also no significant interaction found between learning style preference and instructional condition. This provides further evidence that the failure to find significant support for the meshing hypothesis in Research Question 1 was not likely due to differences in listening versus reading test difficulty.

A second limitation of the study discussed for Research Question 1 pertained to the fact that regardless of mode of instruction, comprehension was assessed using written questions only. The same limitation also applies to Research Question 2. We considered that holding the format of the assessment constant would allow only one variable (in this case, mode of instruction) to be varied within the study. A written format was chosen over a listening format because this is consistent with how most tests are given. However, it could be argued that using the same (written) format for the assessment of learning may have favored those individuals who had a stronger visual learning style preference and, thus, masked evidence supporting the meshing hypothesis. Indeed, it was found in both Research Questions 1 and 2 that participants with a visual word learning style preference performed significantly better than those with an auditory learning style preference on both the listening and reading comprehension tests, both of which were assessed by written questions. However, it also should be recalled that both learning style preference groups performed better on the listening aptitude test than the reading aptitude test in Research Question 1, even though both were assessed with written questions. Regardless of these potential limitations to the study design, it should be kept in mind that the critical test of the meshing hypothesis rests in finding a significant *interaction* between learning style preference and either aptitude (Research Question 1) or learning based on mode of instruction (Research Question 2). This was not found in either case. Nonetheless, it may be important in future studies to determine if the meshing hypothesis may be supported if both modes of instruction as well as assessment measures of aptitude or learning are given in both a listening and a written format.

Participants in this study were college-educated adults, and therefore, the results can only be generalized to similar populations with well-developed listening and reading comprehension skills. It will be particularly important for future research to repeat this same study with children of different ages who are in the process of developing reading skills to determine the extent to which mode of instruction, learning aptitude, and learning style preference may affect individual differences in learning outcomes at different stages during the development of language and literacy skills. It would also be important to determine longitudinally the extent to which mode of instruction or learning styles influence literacy

outcomes when instruction is provided over a longer period of time.

This research focused narrowly on verbal comprehension skills and the extent to which learning differed when instruction is presented via an audiobook compared with e-text. While there are many different schemes for classifying individual learning styles, we used only one learning style inventory (the Rundle and Dunn Building Excellence Inventory) and within that inventory focused only on auditory and visual word learning styles. Thus, the degree to which the results of this study generalize to other disciplines or other learning styles cannot be established by this study. Furthermore, instruction used in this study was given only one time and relied on participants learning information from the preface and one chapter in a nonfiction book. The extent to which the results of this study can be generalized to other forms of instruction, longer durations of instruction, and other types of material cannot be established.

In Research Question 2, the sample size was substantially reduced by the random placement of participants into different instructional conditions (audiobook, e-text) and because of the categorical analyses. Therefore, for several of the analyses concerned with finding relations among individual differences in learning style preferences or aptitudes and mode of instruction, the lack of statistical significance may be influenced by a lack of power due to a modest sample size. Nonetheless, when the results from both Research Question 1, which included a much larger sample size ($N = 121$) and Research Question 2 ($n = 61$) are considered in their entirety, they consistently fail to provide any empirical evidence that suggests individuals will learn significantly better when they are provided instruction in a mode that meshes with their preferred or stronger learning style than in a mode that does not.

Conclusion

The American education system as well as the general public has come to believe that optimal learning occurs if individuals are taught in their preferred learning style. Dekker et al. (2012) surveyed 242 primary and secondary school teachers from the United Kingdom ($n = 137$) and the Netherlands ($n = 105$) who were enthusiastic about applying neuroscientific findings into their instruction. It was assumed that this population, given their high level of interest, would be current on effective research-based practices. The participants were given statements and were asked if the statements were "correct," "incorrect," or "do not know." Results showed that 93% of teachers from the United Kingdom and 96% of teachers from the Netherlands answered "correct" to the statement: "Individuals learn better when they receive information in their preferred learning style (e.g., auditory, visual, kinesthetic)." The results of this study demonstrate how pervasive the misinformation of learning styles is in everyday classroom practice around the world.

The idea of teaching to an individual's learning style is attractive. According to learning styles theory, if an individual is struggling to learn new material, it is possible that his or her poor performance results from not being taught in a mode that meshes with the individual's preferred learning style. Thus, educators and professional development leaders spend time and resources assessing their students' learning style and developing instruction to

specifically match a student's preferred learning styles. It is common for lesson plans to include a section in which teachers are asked to explain how they will accommodate the different learning styles of students in their classroom. Therefore, the findings from this study have considerable relevance for educational theory and practice.

The main finding from Research Questions 1 and 2 that may have a substantial influence on current educational practice is that when participants were categorized by their preferred learning style, either auditory or visual word, those who were classified as visual word learners performed better, compared with auditory learners, on verbal comprehension measures. In other words, visual word learners scored higher than auditory learners on both the reading and the listening aptitude tests and the *Unbroken* comprehension tests. Therefore, and counter to current educational beliefs and practices, educators may actually be doing a disservice to auditory learners by continually accommodating their auditory learning style preference by providing them instruction that meshes with their auditory learning style, rather than focusing on strengthening their visual word skills. It is important to keep in mind that most testing, from state standardized education assessments to college admission tests, is presented in a written word format only. Thus, it is important to give students as much experience with written material as possible to help them build these skills, regardless of their preferred learning style. Rather than continually accommodating auditory learners' preference with increased instruction in an auditory format, auditory learners might benefit more from receiving instruction that specifically targets and strengthens their visual word skills.

In a review of the learning styles literature, Pashler et al. (2008) did not find empirical support to justify matching instruction to learning style. He and his collaborators brought to light several pressing concerns. First, too often individuals allow their intuitions to shape their beliefs. We base our educational practice on trial and error, or we are complicit in always doing what has always been done. Changing the minds of teachers and teacher educators with regards to learning styles is no small feat. Pashler et al. (2008) stated, "If education is to be transformed into an evidence-based field, it is important not only to identify teaching techniques that have experimental support but also to identify widely held beliefs that affect the choices made by educational practitioners that lack empirical support" (p. 117). The goal of this study was to provide more empirical evidence to guide educational practitioners in making sound judgments pertaining to whether their students will or will not benefit from receiving instruction that meshes with their preferred learning style or aptitude. In the current study, we failed to find any statistically significant, empirical support for tailoring instructional methods to an individual's learning style.

References

Cassidy, S. (2004). Learning styles: An overview of theories, models, and measures. *Educational Psychology*, 24, 419–444. doi.<http://dx.doi.org/10.1080/0144341042000228834>.

Coffield, F., Moseley, D., Hall, E., & Ecclestone, K. (2004). *Learning styles and pedagogy in post-16 learning: A systematic and critical review*. London, England: Learning & Skills Research Centre. Retrieved from <http://lerenleren.nu/bronnen/Learning%20styles%20by%20Coffield%20e.a.pdf>

Dekker, S., Lee, N. C., Howard-Jones, P., & Jolles, J. (2012). Neuromyths in education: Prevalence and predictors of misconceptions among teach-

ers. *Frontiers in Psychology: Educational Psychology*, 429, 1–8. doi: 10.3389/fpsyg.2012.00429

Dunn, R., Dunn, K., & Price, G. E. (1989). *Learning Styles Inventory*. Lawrence, KS: Price Systems.

Gregorc, A. F. (1982). *Gregorc Style Delineator: Development, technical, and administration manual*. Maynard, MA: Gabriel Systems.

Herrmann, N. (1996). *The whole brain business book*. New York, NY: McGraw-Hill.

Kolb, D. (1985). *Learning Style Inventory*. Boston, MA: McBer.

Kozhevnikov, M. (2007). Cognitive styles in the context of modern psychology: Toward an integrated framework of cognitive style. *Psychological Bulletin*, 133, 464–481. doi:10.1037/0033-2909.133.3.464

Paehler, H., McDaniel, M., Rohrer, D., & Bjork, R. (2008). Learning styles: Concepts and evidence. *Psychological Science in the Public Interest*, 9, 105–119. Retrieved from <http://psi.sagepub.com/content/9/3/105.abstract>

Rundle, S., & Dunn, R. (2010). *Learning styles: Online learning style assessments and community*. Retrieved from <http://www.learningstyles.net>

Sternberg, R. J., Grigorenko, E. L., & Zhang, L. (2008). Styles of learning and thinking matter in instruction and assessment. *Perspectives on Psychological Science*, 3, 486–506. doi:10.1111/j.1745-6924.2008.00095.x

Weiderholt, J. L., & Bryant, B. R. (2000). *Gray Oral Reading Test* (4th ed.). Austin, TX: Pro-Ed.

Received August 21, 2013

Revision received June 11, 2014

Accepted June 15, 2014 ■

Members of Underrepresented Groups: Reviewers for Journal Manuscripts Wanted

If you are interested in reviewing manuscripts for APA journals, the APA Publications and Communications Board would like to invite your participation. Manuscript reviewers are vital to the publications process. As a reviewer, you would gain valuable experience in publishing. The P&C Board is particularly interested in encouraging members of underrepresented groups to participate more in this process.

If you are interested in reviewing manuscripts, please write APA Journals at Reviewers@apa.org. Please note the following important points:

- To be selected as a reviewer, you must have published articles in peer-reviewed journals. The experience of publishing provides a reviewer with the basis for preparing a thorough, objective review.
- To be selected, it is critical to be a regular reader of the five to six empirical journals that are most central to the area or journal for which you would like to review. Current knowledge of recently published research provides a reviewer with the knowledge base to evaluate a new submission within the context of existing research.
- To select the appropriate reviewers for each manuscript, the editor needs detailed information. Please include with your letter your vita. In the letter, please identify which APA journal(s) you are interested in, and describe your area of expertise. Be as specific as possible. For example, “social psychology” is not sufficient—you would need to specify “social cognition” or “attitude change” as well.
- Reviewing a manuscript takes time (1–4 hours per manuscript reviewed). If you are selected to review a manuscript, be prepared to invest the necessary time to evaluate the manuscript thoroughly.

APA now has an online video course that provides guidance in reviewing manuscripts. To learn more about the course and to access the video, visit <http://www.apa.org/pubs/authors/review-manuscript-ce-video.aspx>.

Toward an Understanding of Dimensions, Predictors, and the Gender Gap in Written Composition

Young-Suk Kim

Florida State University and Florida Center
for Reading Research

Jeanne Wanzek

Florida State University and Florida Center
for Reading Research

Stephanie Al Otaiba

Southern Methodist University

Brandy Gatlin

Florida State University

We had 3 aims in the present study: (a) to examine the dimensionality of various evaluative approaches to scoring writing samples (e.g., quality, productivity, and curriculum-based measurement [CBM] writing scoring), (b) to investigate unique language and cognitive predictors of the identified dimensions, and (c) to examine gender gap in the identified dimensions of writing. These questions were addressed using data from 2nd- and 3rd-grade students ($N = 494$). Data were analyzed using confirmatory factor analysis and multilevel modeling. Results showed that writing quality, productivity, and CBM scoring were dissociable constructs but that writing quality and CBM scoring were highly related ($r = .82$). Language and cognitive predictors differed among the writing outcomes. Boys had lower writing scores than girls even after accounting for language, reading, attention, spelling, handwriting automaticity, and rapid automatized naming. Results are discussed in light of writing evaluation and a developmental model of writing.

Keywords: dimensionality, writing quality, writing productivity, CBM, gender

Students' writing skill is assessed in multiple ways. To assess a discourse-level writing skill (e.g., ability to write in paragraphs), students are typically asked to write written compositions, and written compositions are evaluated using multiple approaches such as writing quality, writing productivity, or curriculum-based measurement (CBM) writing scoring. Another widely used writing assessment measures a sentence-level writing ability by asking students to produce grammatically correct sentences within a specified time (Writing Fluency task of the Woodcock–Johnson Tests of Achievement–III [WJ–III], Woodcock, McGrew, & Mather, 2001). Despite the existence of various ways of assessing students' writing skill, researchers and practitioners have a limited understanding of how these various assessments and evaluative ap-

proaches are related and whether they tap into or capture similar or dissociable dimensions of writing. A clearer understanding of assessment approaches is needed to advance theories of development and to guide practitioners in using assessment data to inform instruction and intervention. In the present study, we addressed this question with three goals. First, we examined how various approaches to writing assessments converge or diverge into different dimensions, using various evaluative approaches such as writing quality, productivity, and CBM scoring as well as using a widely used sentence-level task, the WJ–III Writing Fluency task. Second, we further examined how language and cognitive skills relate to the identified dimensions. Finally, given the consistent achievement gaps between boys and girls on national writing assessments (e.g., Persky, Dane, & Jin, 2003), we also sought to examine gender differences across the identified dimensions of writing.

This article was published Online First July 7, 2014.

Young-Suk Kim, College of Education and Florida Center for Reading Research, Florida State University; Stephanie Al Otaiba, College of Education, Southern Methodist University; Jeanne Wanzek, College of Education and Florida Center for Reading Research, Florida State University; Brandy Gatlin, College of Education, Florida State University.

This research was supported by Grant P50HD052120 from the National Institute of Child Health and Human Development. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute of Child Health and Human Development. The authors thank study participants including students, teachers, school personnel, and parents.

Correspondence concerning this article should be addressed to Young-Suk Kim, Florida Center for Reading Research, Florida State University, 1114 West Call Street, Tallahassee, FL 32306. E-mail: ykim@fcrr.org

Approaches to Writing Evaluation

According to the simple view of writing (Juel, Griffith, & Gough, 1986), two necessary components of writing are ideation (i.e., generation and organization of ideas) and transcription skills. The first component, ideation, refers to the quality of ideas represented in writing, which is an essential, and arguably the most important, aspect to be evaluated in written compositions. Not surprisingly, writing quality has long and widely been examined in previous studies. Two key indicators of writing quality appear to be the extent of development and organization of ideas (Bereiter & Scardamalia, 1987; Juel et al., 1986). In fact, idea development and organization have been widely examined as indicators of writing

quality in previous studies (Graham, Harris, & Chorzempa, 2002; Graham, Harris, & Mason, 2005; Kim, Al Otaiba, Folsom, & Greulich, 2013; Kim et al., 2011; Olinghouse, 2008). Other widely used assessments of writing examine similar aspects. For example, the fourth edition of the Test of Written Language (TOWL-4; Hammill & Page, 2009) includes theme development and organization, and another widely used writing evaluation approach in U.S. schools (Gansle et al., 2006), the 6 + 1 Trait Rubric (Northwest Regional Laboratory, 2011), includes idea development and organizational or structural aspects in addition to other aspects such as word choice, sentence fluency, voice, presentation, and conventions.

The other component of the simple view of writing, transcription skill, allows generated ideas to be produced in written text and facilitates idea generation and development (Berninger et al., 1997; Graham, Berninger, Abbott, Abbott, & Whittaker, 1997; Graham, Harris, & Fink, 2000; Kim et al., 2011). Therefore, the amount of written composition is constrained by transcription skills to a large extent, particularly for beginning writers. Not surprisingly, writing productivity is another widely examined dimension of writing (e.g., Abbott & Berninger, 1993; Berman & Verhoeven, 2002; Kim et al., 2011, 2014; Mackie & Dockrell, 2004; Olinghouse & Graham, 2009; Scott & Windsor, 2000). Note that although the term *writing fluency* has been used often to refer to a similar construct, we use the term *writing productivity*, because we are specifically referring to the amount of text produced, not the automaticity, effortlessness, or coordination of multiple processes, which are defining characteristics of fluency (Berninger et al., 2010; LaBerge & Samuels, 1974). In addition, writing fluency has been conceptualized to refer to CBM writing (Ritchey et al., in press). Although the amount of text alone is not generally considered a yardstick or goal of good writing, good written composition requires a certain amount of text for the ideas to be sufficiently developed and articulated. Previous studies have examined writing productivity, and it has been shown to be a dissociable dimension from writing quality (Kim, Al Otaiba, Sidler, Greulich, & Puranik, 2014; Wagner et al., 2011), and correlations between writing quality and productivity tend to be fairly strong for children in the elementary grades (e.g., $.65 \leq rs \leq .82$; Abbott & Berninger, 1993; Kim et al., 2014; Olinghouse & Graham, 2009). Writing productivity is measured using various indicators such as the total number of words, number of ideas, number of different words, and/or number of sentences (Kim et al., 2014; Kim, Park, & Park, 2013; Puranik, Lombardino, & Altmann, 2008; Wagner et al., 2011).

A third evaluative approach to writing employed in the present study is CBM scoring. CBM writing scoring includes some unique evaluative tools not included in the writing quality and productivity indicators noted previously. Along with reading and math CBM measures, CBM writing measures are considered global outcome measures, or indicators, of students' overall writing performance (Deno, 1985) that are intended to signal whether the student needs further diagnosis and intervention. CBM writing measures were initially developed to screen and monitor progress in writing skills for students at risk for writing difficulty. Students are typically asked to write for 3–5 min in response to prompts (Coker & Ritchey, 2010; McMaster, Du, & Pétursdóttir, 2009; McMaster et al., 2011), and their writing is evaluated using various scoring tools such as number of words written, correct word sequences (two

adjacent words that are grammatically correct and spelled correctly), incorrect word sequences, words spelled correctly, percentage of correct word sequences, and correct minus incorrect word sequences (see Graham, Harris, & Herbert, 2011; McMaster & Espin, 2007, for a review). Note that number of words written is not unique to the CBM writing scoring as it has been used as an indicator of writing productivity.

CBM writing measures have been shown to be reliable, and students' scores on CBM writing tend to be related to other writing measures with validity coefficients in the moderate range (see Graham et al., 2011, and McMaster & Espin, 2007, for reviews; Lembke, Deno, & Hall, 2003; McMaster et al., 2009). In particular, the correct minus incorrect word sequences (CIWS) score tends to be the most strongly related to other writing measures with coefficients ranging from .60 to .75 (Espin et al., 2000; Espin, Weissenburger, & Benson, 2004). Recently, the percentage of correct word sequences (%CWS), along with the CIWS, has also been shown to be highly ($r = .61$) related to a normed writing task (Test of Written Language, 3rd ed. 1996) for children in middle school (Amato & Watkins, 2011).

Despite the reliability and validity evidence for CBM writing scoring procedures described in these previous studies, it is not clear how CBM writing scores should be conceptualized in terms of dimensionality. That is, do CBM writing scores capture dimensions such as writing quality or writing productivity, or do they measure a separate, overall global outcome measure of writing? Recently, CBM writing measures have been described as *writing fluency*, which is defined as the ease with which an individual "produces written text" and includes both "*text generation* (translating ideas into words, sentences, paragraphs, and so on) and *transcription* (translating words, sentences, and higher levels of discourse into print)." (italics in the original text; Ritchey et al., in press). A critical question is whether potential writing fluency indicators capture a dissociable dimension, apart from other widely examined dimensions such as writing quality and productivity. Although the theoretical foundations for using CBM writing scores as measurement are still in their nascent stage, we included CBM in the present study because of validity evidence with other writing measures, and its potential practical utility for progress-monitoring purposes, as CBM indicators have been shown to be sensitive to growth over time within a short time period (e.g., 2 weeks; see Espin et al., 2004; McMaster & Espin, 2007).

Finally, although writing skill is typically assessed by asking the child to produce a written composition, other tasks also have been used. One such widely used standardized subtest is the Writing Fluency task of the WJ-III (Woodcock et al., 2001). This task assesses sentence-level, rather than paragraph-level, writing. Children are presented with a picture and three words, and they are asked to write a sentence about the picture using the three words. The child's score is the number of correct and meaningful written sentences based on the three words that were presented. However, how the WJ-III Writing Fluency relates to other dimensions of writing is an open question.

In the present study, we examined dimensionality of writing using children's data from written compositions as well as the Writing Fluency task of the WJ-III. Children's written compositions were evaluated by indicators of writing quality, productivity, and CBM writing scores. For the Writing Fluency task of the WJ-III, scores following the WJ-III scoring guidelines were used.

Our goal in the present study was to extend our understanding of writing dimensionality. Previous studies have shown that writing quality, productivity, spelling and writing conventions, and syntactic complexity are dissociable dimensions for typically developing children in Grades 1 and 4, as well as children with language impairments (Kim et al., 2014; Puranik et al., 2008; Wagner et al., 2011). In the present study, we expand this line of research by examining how CBM scores and the Writing Fluency task of the WJ-III are related to writing quality and productivity dimensions using data from children in Grades 2 and 3.

Predictors of Writing Skills

As noted earlier, writing is composed of at least two component skills: transcription skills and ideation (Berninger & Swanson, 1994; Juel et al., 1986). Transcription skills such as spelling and handwriting allow mental resources such as attention and working memory to be available for idea generation and translation processes (Berninger & Swanson, 1994; Graham, 1990; Graham et al., 1997, 2000; Scardamalia, Bereiter, & Goleman, 1982). Much evidence supports the role of transcription skills in writing (Berninger, 1999; Graham et al., 1997; Jones & Christensen, 1999; Kim et al., 2011, in press; Wagner et al., 2011). Handwriting skill is typically assessed by asking the child to write alphabet letters or to copy sentences or paragraphs as accurately and quickly as possible within a specified time (e.g., Abbott & Berninger, 1993; Graham et al., 1997; Kim et al., 2011; Wagner et al., 2011).

Although ideation, the other component of writing according to the simple view of writing, is challenging to directly measure, it has been largely measured by means of oral language use (e.g., Chenoweth & Hayes, 2003; Hayes, 2012). Generated ideas cannot be produced without being translated into oral language because the child has to express ideas using appropriate words, encode them using appropriate syntactic structure, and organize and present them in a logical sequence. Therefore, oral language proficiency would determine how the generated ideas are adequately expressed. Evidence of the importance of oral language in written composition is accumulating from beginning writers to those in middle school (Berninger & Abbott, 2010; Kent, Wanzek, Petscher, Al Otaiba, & Kim, in press; Kim et al., 2011, 2013, 2014; Olinghouse, 2008) as well as children with language impairment (Dockrell, Lindsay, & Connelly, 2009; Dockrell & Connelly, in press; Kim, Puranik, & Al Otaiba, in press; Puranik, Lombardino, & Altmann, 2007). Given that writing is a production or constructed-response task, children's transcription skills constrain the extent to which generated ideas can be transcribed into generated text (Berninger, Abbott, Abbott, Graham, & Richards, 2002; Juel et al., 1986).

In addition to these previously noted skills, the not-so-simple view of writing states that executive function processes such as attention, planning, self-regulation, and working memory are critical supports for writing development (Berninger & Winn, 2006). Attention, in particular, has been shown to be related to writing for children in first and second grade (Hooper et al., 2011; Hooper, Swartz, Wakely, de Kruif, & Montgomery, 2002; Kent et al., in press; Kim, Al Otaiba, et al., 2014). Additional evidence underscoring the importance of attention in writing comes from studies with children who have attention deficits or attention-deficit/hyperactivity disorder (ADHD); converging evidence suggests that

students with ADHD made more spelling and grammatical errors (Casas, Ferrer, & Fortea, 2013; Gregg, Coleman, Stennett, & Davis, 2002; Re, Pedron, & Cornoldi, 2007), made more content errors or digressions and demonstrated weaker text structure features than children without ADHD (Casas et al., 2013).

Individual differences in reading also have been shown to matter for children's writing development (Shanahan, 2006). Studies have shown that reading comprehension was related to written composition quality and productivity for children in elementary and middle school grades (Berninger & Abbott, 2010; Berninger et al., 2002; Kim, Al Otaiba, et al., 2013, 2014). Children's reading ability might influence written composition skill via reading experiences. Greater reading ability and consequent text reading might allow the opportunity for the child to acquire vocabulary and syntactic structures, and organization of written text as well as content (Berninger et al., 2006). In fact, children with impaired reading comprehension had weaker story content and organization in their writing (Cragg & Nation, 2006).

Writing involves juggling of multiple processes to even greater extent than in reading. Therefore, the ability to coordinate multiple aspects is likely to be important. Some previous studies have examined rapid automatized naming (RAN) in this regard as a potential predictor of writing. Numerous studies have shown that rapid automatized naming is related to reading (Compton, DeFries, & Olson, 2001; de Jong, & van der Leij, 2003; Kim, 2011; Kirby, Parrila, & Pfeiffer, 2003; Savage et al., 2005; Wolf & Bowers, 1999; Wolf & O'Brien, 2001). However, despite a robust relation to reading in various languages, researchers differ about what it exactly measures, and hypotheses include phonological processing (Wagner & Torgesen, 1987), automaticity of processes (Bowers, 1995; LaBerge & Samuels, 1974; Spring & Davis, 1988), global processing speed (Kail & Hall, 1994), and multiple constructs such as lexical access, automaticity, attentional, visual, and articulatory processes (Wolf & Bowers, 1999). If RAN measures automaticity of processes, its influence might largely overlap with that of automaticity of transcription skills and thus may not be related to writing over and above transcription skills. In contrast, if RAN captures multiple constructs beyond what is captured by transcription skills, it would be related to writing over and above transcription skills. Although RAN has not been examined for young English-speaking children, there is some emerging evidence from studies with Chinese children that suggests that RAN is related to writing (Chan, Ho, Tsang, Lee, & Chung, 2006; Ding, Richman, Yang, & Guo, 2010; but see Yan et al., 2012).

Gender and Writing

Gender appears to matter in children's writing achievement. Girls have outperformed boys in writing consistently across grades ever since writing was included in the National Assessment of Educational Progress (National Center for Education Statistics, 2011). For instance, in 2002, in which writing was assessed in children in Grade 4 as well as those in Grades 8 and 12, girls outperformed boys in all the three grades with gaps ranging from 17 to 25 points (National Center for Education Statistics, 2003). Similarly, gender gaps have been reported for children in elementary grades (Berninger & Fuller, 1992; Knudson, 1995). Despite these consistent gender gaps in writing, our understanding about gender gaps in writing and potential sources of gender gaps is

limited, particularly for children in elementary grades. One potential source of gender gaps seen in older students is their attitude toward writing. Among adolescents, males tend to have less positive attitudes toward writing than do females (Knudson, 1992; Pajares & Valiante, 1999), and see less value in writing and express less satisfaction with writing activities (Lee, 2013). Studies of younger students have reported mixed findings about the relation of attitude toward writing and children's writing skill. Knudson (1995) investigated gender and writing attitude with children in Grades 2 and 6 and found that children's attitude toward writing predicted their writing skill. In contrast, a study with children in Grades 1 and 3 revealed that girls had more positive attitudes than boys toward writing as early as in Grade 1, but this difference was not related to their writing skill (Graham, Berninger, & Fan, 2007).

Another potential source of gender gaps in writing achievement is reading or reading-related skills. As noted earlier, evidence suggests that reading is one of the component skills of writing. Evidence also indicates that male students have been consistently outperformed by female students in reading (e.g., National Center for Education Statistics, 2011), and a greater number of boys are identified with reading disabilities (Hawke, Olson, Willcutt, Wadsworth, & DeFries, 2009; Miles, Haslum, & Wheeler, 1998; Yoshimasu et al., 2010; but see Shaywitz, Shaywitz, Fletcher, & Escobar, 1990). Therefore, differences in reading or reading-related skills might explain differences in writing skills between boys and girls. Furthermore, boys in Grades 1, 2, and 3 had lower scores in another writing component skill, transcription skill (Berninger & Fuller, 1992). In the present study, we examined whether gender differences were found for children in Grades 2 and 3 in the identified writing dimensions, and if so, to what extent gender differences were explained by the included language and cognitive skills (e.g., reading, attention, and transcription skills).

Present Study

The primary goal of the present study was to examine the dimensionality of various writing evaluation approaches, predictors of various dimensions, and the gender gap in writing. Specific research questions were as follows.

1. What are the relations of CBM writing measures (i.e., CIWS and %CWS) and the WJ-III Writing Fluency task to writing quality and writing productivity indicators? Are CBM writing measures and the WJ-III Writing Fluency task measure dissociable dimensions from writing quality and writing productivity?
2. How are language and cognitive skills related to the identified writing dimensions?
3. Are there performance differences between boys and girls in the identified writing dimensions (e.g., writing quality and productivity) after accounting for children's language and cognitive skills?

To address these questions, we used data from second- and third-grade children ($N = 494$) who were administered multiple writing tasks: written compositions in response to three prompts (one normed task and two experimental tasks) and a sentence-level task, the WJ-III Writing Fluency task. Students' compositions were evaluated using a variety of approaches including writing quality indicators such as idea development and organization, writing productivity indicators such as number of words written

and number of ideas, CBM writing scoring such as CIWS and %CWS, and scoring protocols in the standardized tasks. Language and cognitive skills included oral language, reading, transcription (spelling and handwriting fluency), attention, and rapid automated naming.

We hypothesized that writing quality and productivity would be dissociable dimensions based on previous studies (Kim, Al Otaiba, et al., 2014; Puranik et al., 2008; Wagner et al., 2011). We also hypothesized that the CBM writing scores would be a dissociable construct because although validity coefficients of CIWS and %CWS were acceptable, they are not extremely highly correlated with other writing measures (e.g., Amato & Watkins, 2011; McMaster & Espin, 2007). In contrast, we did not have a priori prediction about the WJ-III Writing Fluency task. It was also hypothesized that various language and cognitive skills would be differentially related to different writing outcomes based on a prior study (Kim, Al Otaiba, et al., 2014); Finally, gender differences were hypothesized, and language and literacy skills were expected to explain gender differences to some extent.

Method

Participants and Sites

Students in the present study included 494 children in Grades 2 (mean age = 7.80 years) and 3 (mean age = 8.82 years). These students were drawn from 76 classrooms in 10 schools in a midsized city. The students were 51.2% male, and 76.1% received free or reduced-price lunch. Six of the 10 schools were Title I schools, indicating that the majority of the students in the school were eligible for the free or reduced-price lunch program. Students' racial backgrounds were as follows: 60% African Americans, 29% Whites, and the rest Asians and multiracial children. The students and their families had consented for their participation, and all guidelines for human research protection continued to be followed in the present study.

Measures

Writing tasks. Four tasks were used to assess children's written composition skill: two standardized and normed tasks, and two experimental prompts. The first task was the Writing Fluency subtest of the Woodcock-Johnson Tests of Achievement (3rd ed., or WJ-III; Woodcock et al., 2001). In this subtest, students were provided with a series of pictures and three corresponding words and were instructed to write a sentence about the picture that included the words given. Students were given 7 min to complete as many sentences as they could. For the scoring of this subtest, we used standard scoring procedures outlined in the testing manual. Namely, students received 1 point for each complete sentence. In order to receive credit, the sentence had to be clear in meaning and include critical words to make the sentence reasonable. Students were not penalized for errors in punctuation, spelling, or capitalization, or for poor handwriting. Using the Rasch analysis procedure, the reliability coefficient was reported to be .72 for 7- and 8-year-olds (McGrew, Schrank, & Woodcock, 2007).

We also asked children to write on three prompts: one prompt from the Written Essay test of the Wechsler Individual Achievement Test (3rd ed., or WIAT-III; Wechsler, 2009) and two exper-

imental prompts, one narrative prompt and one expository prompt. The WIAT-III was selected as a widely used writing assessment that could be compared with other research (e.g., see Berninger & Abbott, 2010). In the WIAT-III Essay task, children were asked to write about a favorite game and include at least three reasons as support. Note that standard scores in this task are available starting with children in Grade 3 and not available for children in Grade 2. Despite lack of standard scores, this task was deemed useful for children in Grade 2 for the purpose of examining dimensionality and predictive relations. In addition, assessors confirmed that this topic did not appear to be difficult for children in Grade 2.

The experimental narrative prompt was "One day when I got home from school . . ." Children were asked to write about any interesting events that occurred responding to the prompt (Kim et al., 2013, Kim, Al Otaiba, et al., 2014; McMaster et al., 2009; 2011). The experimental expository prompt was adapted from a previous study (Wagner et al., 2011). In this task, children were asked to write about a classroom pet they would like and explain why. For each prompt, children were given a 10-min time limit.

Writing Evaluation

Children's written compositions for the WIAT Essay Composition task and two experimental prompts were evaluated on writing quality, writing productivity, and CBM writing measure scoring (discussed later). In addition, the WIAT essay was scored according to the examiner's manual. Children's responses to the WJ-III Writing Fluency task were evaluated only according to the examiner's manual previously noted because the responses were sentences, not passage-level composition.

Writing quality scoring. The quality of children's written composition was evaluated on the extent to which their ideas were developed and the extent to which the ideas were presented in an organized manner, on a rating scale of 1 to 7. In this idea development aspect, high scores were given to compositions with rich and detailed ideas and ideas with unique and interesting perspectives. In the organization aspect, compositions with logical sequence and transitions of expressed ideas with overall structure of beginning, middle, and end received high scores. These were similar to the 6-point scale version of the 6 + 1 Trait Rubric but were adapted to a 1-7 rating scale, representing low quality and high quality, respectively. When using 45 writing samples per prompt, reliabilities (Cohen's kappa) ranged from .82 to .88 for ideas and organization.

Writing productivity scoring. Two indicators were used for writing productivity: total number of words written and number of ideas. The number of words has been widely used as an indicator of compositional productivity in writing (e.g., Abbott & Berninger, 1993; Berman & Verhoeven, 2002; Kim et al., 2011; Mackie & Dockrell, 2004; Puranik et al., 2008; Scott & Windsor, 2000; Wagner et al., 2011). Words were defined as real words recognizable in the context of the child's writing despite some spelling errors. Random strings of letters or sequences were not counted as words. Random strings of letters were identified by comparing a record of what the child said she had written to her written composition. These were extremely rare in the sample (less than 10). The number of ideas was a total number of propositions, which were defined as predicate and argument. For example, "I went upstairs and took a bath" was counted as two ideas (see, e.g.,

Kim et al., 2011, 2013; Puranik et al., 2008). Repeated ideas were only counted once. When using 45 writing samples per prompt, reliabilities were .88 for the number of ideas (kappa) and .99 for the number of words (similarity).

Curriculum-based measure scoring. Each essay was individually analyzed for curriculum-based measures (CBM) including the *correct word sequence* ("any two adjacent, correctly spelled words that are acceptable within the context of the sample"; McMaster & Espin, 2007, p. 76), and the *incorrect word sequence* ("any two adjacent letters that are incorrect"; McMaster & Espin, 2007, p. 76). From these, a correct minus incorrect word sequence (CIWS) was obtained by subtracting incorrect words from correct word sequence. The percentage of correct word sequences (%CWS) was calculated by dividing the number of CWS by the total number of words written. In the data analysis, we used CIWS and %CWS for two reasons: (a) number of words written has been used as an indicator writing productivity and, thus, is not unique to CBM writing, and correct word sequence is highly related to the number of words written (because children who write more tend to have greater number of correct sequences); and (b) evidence indicates that CIWS and %CWS have greater validity coefficients with other writing tasks than the other CBM writing scoring (e.g., McMaster & Espin, 2007). Reliability for each type of scoring was established using 45 pieces per prompt. We used an equation that produced quotients to indicate the proximity of the coder's score for each measure to that of the primary coder (i.e., similarity coefficients; Shrout & Fleiss, 1979), and reliability for each measure ranged from .92 to .99.

WIAT standardized scoring. In addition to the previously noted evaluative measures, students' compositions for the WIAT Essay Composition task were scored according to the manual. The WIAT scoring includes the total number of words, thematic development, and text organization (theme and organization hereafter), and a supplemental score called the *grammatical score*. The grammatical score is highly similar to CIWS in CBM writing although slight differences are found in operationalization (e.g., WIAT does not give credit for titles or endings such as "The End," whereas conventional CBM writing does). The unique scoring in the WIAT task, thus, is the theme and organization, and students' compositions were assigned scores in the following categories: introduction, conclusion, paragraphs, transitions, reasons why, and elaborations. The maximum score possible for the theme and organization component was 20 points. Interrater reliability was established by having two independent coders score 50 essays and comparing individual points assigned. The number of agreements was divided by the total number of agreements plus disagreements, resulting in a reliability coefficient of .85. A standard score for theme and organization was computed for each student based on his or her chronological age at the time of testing. The standard score for the WIAT Essay Composition task is a composite of the standard score for theme and organization and for total number of words written.

Predictors

Predictors were selected based on our review of the literature and included oral language, reading, spelling, handwriting fluency (letter writing and story copying tasks), attention, and rapid automatized naming.

Oral language. Children's oral language skill was measured by the following three tasks: WJ-III Picture Vocabulary (Woodcock et al., 2001), the Test of Narrative Language Narrative Comprehension subtest (Gillam & Pearson, 2004), and the Oral and Written Language Scales Listening Comprehension subtest (Carrow-Woolfolk, 2011). In the Picture Vocabulary task, children were asked to identify pictured objects. Test-retest reliability is reported as .71-.73 for 7- and 8-year-olds (McGrew et al., 2007). The Narrative Comprehension subtest of the Test of Narrative Language includes three individual tasks in which each student listens to a short story and is then asked to answer specific comprehension questions. The internal consistency of this subtest is .87, and test-retest reliability is .85 (Gillam & Pearson, 2004). In the Listening Comprehension subtest of the Oral and Written Language Scales, students listen to a stimulus sentence and are asked to point to one of four pictures that corresponds to the sentence read aloud by the tester. This subtest's reported split-half internal reliability ranges from .96 to .97 for the age group of our sample (Carrow-Woolfolk, 2011).

Reading. Children's reading skill was assessed using five measures: the WJ-III Letter Word Identification and Passage Comprehension subtests (Woodcock et al., 2001), the Sight Word Efficiency subtest of the Test of Word Reading Efficiency (2nd ed., Torgesen, Wagner, & Rashotte, 2012), the Oral Reading Fluency subtest of the WIAT-III (Wechsler, 2009), and the Test of Silent Reading Efficiency and Comprehension (Wagner, Torgesen, Rashotte, & Pearson, 2010). For the Letter Word Identification task, the child is asked to read aloud letters and words of increasing difficulty. For the WJ-III Passage Comprehension subtest, students are asked to silently read a short passage and provide a missing word that makes sense within the context of the passage. Reliabilities (test-retest) are reported to be .96 for both the Letter Word Identification and the Passage Comprehension subtests for students in the age range of the students we assessed (McGrew et al., 2007). In the Sight Word Efficiency task, the child is asked to read words of increasing difficulty with accuracy and speed. Test-retest reliability for the Sight Word Efficiency is reported to be .93 for 6- and 7-year-olds and .92 for 8- to 12-year-olds. For the WIAT Oral Reading Fluency task, the child is asked to read two grade-level passages aloud. The student is timed during both readings, and the completion time is recorded in seconds for each prompt. Each raw score is then used to compute an average weighted raw score to determine oral reading fluency. Test-retest reliability for the WIAT Oral Reading Fluency subtest is reported as .93. For the Test of Silent Reading Efficiency and Comprehension, the student is given 3 min to read a series of statements and determine if each statement is true or not. The authors report alternate-form reliability coefficients ranging from .87 to .95 for students in Grades 2 and 3.

Spelling. Children's spelling skill was measured by a dictation task, the WJ-III Spelling subtest (Woodcock et al., 2001). Once a student misspells six consecutive words, the test is discontinued. The authors of this assessment report test-retest reliability coefficients of .91 and .88 for 7- and 8-year-olds, respectively.

Letter writing automaticity. The WIAT-III Alphabet Writing Fluency task was used, in which children were asked to write as many letters of the alphabet as possible with accuracy. This task assesses how well children access, retrieve, and write letter forms automatically. Research assistants asked children to write as many

letters of the alphabet as they could in a 30-s time period. Children received a score for the number of correctly written letters. One point was awarded for each correctly formed letter. Interrater reliability (Cohen's kappa) for this subtest was .88 for our sample.

Story copying. Another transcription skill, the ability to copy letters, was measured by an experimental story copying task. In this task, students were instructed to copy a narrative story titled "Can Buster Sleep Inside Tonight?" as fast as they could. The story had 519 words and involves a dog named Buster being muddy and being bathed so that he could sleep inside. Students were given 1 min to write as much of the story verbatim as possible. Children received a score for the number of letters correctly formed, which was calculated as the difference between the number of letters attempted and the number of letter errors made. Interrater reliability (Cohen's kappa) for this measure was established at .91.

Attention. The first nine items of the Strengths and Weaknesses of ADHD Symptoms and Normal Behavior Scale (e.g., SWAN; Swanson et al., 2006) were used to measure children's attentiveness. SWAN is a behavioral checklist that includes 30 items that are rated on a 7-point scale ranging from a score of 1 (*far below average*) to 7 (*far above average*) to allow for ratings of relative strengths (above average) as well as weaknesses (below average). The first nine items are related to sustaining attention on tasks or play activities (e.g., "Engages in tasks that require sustained mental effort") while the other items assess hyperactivity and aggression. A recent study showed that the first nine items indeed captures the respondent's ability to regulate attention (Sáez, Folsom, Al Otaiba, & Schatschneider, 2012). Higher scores represent greater attentiveness. Teachers completed the SWAN checklist in the spring. Cronbach's alpha across the nine items was .91.

Rapid automatized naming. The Letters subtest of the Rapid Automatized Naming (RAN) test (Wolf & Denckla, 2005) was used. For this subtest, each examinee's completion time for naming a series of alternating lowercase letters was recorded. Test-retest reliability is .89 for children in elementary grades (Wolf & Denckla, 2005).

Procedures

All assessments for the current study were conducted during the spring of the school year. Research assistants were trained prior to each assessment round, which consisted of two individual rounds and two small group sessions. Each research assistant spent approximately 2 hr in training and subsequent practice sessions for each round of assessments and was required to pass a fidelity check before administering assessments to the participants in order to ensure accuracy in administration and scoring. The trained research assistants assessed children individually during two sessions; the first session included the Test of Word Reading Efficiency, Test of Narrative Language Narrative Comprehension subtest, RAN, and WIAT Oral Reading Fluency, and the second session included the WJ-III subtests and the Oral and Written Language Scales Listening Comprehension test. We varied the order of assessments within each session across children in order to reduce fatigue effect. Then, all spelling and writing assessments were administered in small groups over two additional sessions. Throughout the assessments, students were given breaks as needed. Trained research assistants scored students' letter writing automaticity, story copying, spelling, and writing, and research assistants were

trained to use each rubric on a small subset of the sample through practice and discussion of scoring issues.

Data Analysis Strategy

Primary analytic strategies were confirmatory factor analysis (CFA) and multilevel modeling. In the latent variable approach (e.g., CFA), common variance among multiple indicators is used for a construct, and thus measurement error is reduced (Bollen, 1989; Kline, 2005). The first research question, dimensionality of writing, was examined using CFA. Assumptions (univariate and multivariate normality) were checked prior to analysis and were met. Model fits were evaluated using the following multiple indices: chi-square, comparative fit index (CFI), Tucker-Lewis index (TLI), root-mean-square error of approximation (RMSEA), and standardized root-mean-square residuals (SRMR). Differences in model fits for two nested models were evaluated by comparing chi-square differences between the two

models. Confirmatory factor analysis was conducted using Mplus 7 (Muthén & Muthén, 2012). Because children were nested within classroom and schools, the Research Questions 2 and 3 were addressed using three-level multilevel modeling. PROC MIXED procedure of SAS 9.3 was used. Factor scores from CFA models (e.g., scores in the identified writing dimensions) were used in the multilevel modeling.

Results

Descriptive Statistics and Factor Analysis

Table 1 shows the means and standard deviations of writing scores by grade and gender. Where available, standard scores are presented. Note that the WIAT writing composition task was not normed for children in second grade, and thus, standard scores are

Table 1
Means (Standard Deviations) of Writing Measures

Variable	Grade 3			Grade 2			Loadings
	Entire sample	Males	Females	Entire sample	Males	Females	
WIAT total raw score ^a	88.82 (35.23)	81.90 (32.90)	97.55 (36.25)	79.89 (36.09)	71.35 (36.35)	87.21 (34.35)	NA
WIAT total score: SS	107.92 (13.74)	105.73 (14.06)	110.68 (12.87)	NA	NA	NA	NA
WIAT theme & organization raw	6.58 (2.84)	6.28 (2.90)	6.97 (2.74)	5.62 (2.59)	5.38 (2.65)	5.83 (2.53)	.72
WIAT theme & organization SS	105.09 (16.02)	103.71 (16.53)	106.83 (15.24)	NA	NA	NA	NA
WJ-III Writing Fluency raw	13.62 (4.41)	12.89 (4.11)	14.55 (4.62)	10.11 (4.97)	9.48 (4.92)	10.68 (4.96)	.67
WJ-III Writing Fluency SS	98.24 (15.78)	96.34 (13.85)	100.63 (17.69)	95.09 (26.54)	93.44 (24.29)	96.57 (28.44)	NA
Writing quality indicators							
WIAT idea quality	3.89 (0.88)	3.81 (0.83)	3.98 (0.93)	3.44 (0.76)	3.32 (0.81)	3.55 (0.71)	.66
WIAT organization	3.25 (0.89)	3.21 (0.86)	3.30 (0.94)	2.88 (0.82)	2.79 (0.86)	2.96 (0.79)	.70
Narrative idea quality	4.46 (1.00)	4.30 (0.92)	4.66 (1.06)	4.10 (1.10)	3.99 (1.17)	4.19 (1.04)	.65
Narrative organization	3.56 (0.87)	3.44 (0.76)	3.71 (0.97)	3.16 (0.78)	3.10 (0.84)	3.21 (0.73)	.63
Pet idea quality	3.76 (0.80)	3.66 (0.76)	3.88 (0.83)	3.55 (0.81)	3.39 (0.80)	3.70 (0.80)	.54
Pet organization	2.96 (0.69)	2.92 (0.71)	3.02 (0.67)	2.66 (0.69)	2.53 (0.65)	2.77 (0.69)	.60
CBM scores							
WIAT CWS	63.53 (31.85)	57.40 (30.30)	71.19 (32.27)	52.93 (29.97)	45.40 (28.15)	59.23 (20.11)	NA
WIAT IWS	26.43 (16.10)	25.47 (14.16)	27.71 (18.22)	29.55 (20.52)	27.94 (22.24)	30.91 (18.93)	NA
Narrative CWS	66.35 (35.35)	59.94 (31.00)	74.63 (38.89)	54.68 (28.01)	48.50 (25.65)	30.91 (18.93)	NA
Narrative IWS	32.36 (18.33)	31.63 (17.25)	33.30 (19.68)	37.19 (23.50)	34.82 (25.38)	39.16 (21.73)	NA
Pet CWS	62.87 (34.33)	54.31 (30.10)	73.22 (36.34)	54.10 (33.52)	45.25 (29.47)	61.69 (35.01)	NA
Pet IWS	25.57 (17.83)	24.91 (18.32)	26.36 (17.27)	27.40 (21.29)	26.25 (21.29)	28.39 (21.28)	NA
WIAT %CWS	76 (18)	74 (18)	78 (18)	69 (22)	68 (22)	71 (22)	.87
Narrative %CWS	73 (17)	72 (17)	74 (18)	66 (20)	66 (20)	66 (20)	.80
Pet %CWS	78 (20)	75 (21)	81 (19)	72 (22)	69 (22)	74 (22)	.78
WIAT CIWS	37.09 (35.02)	31.99 (33.18)	43.48 (36.35)	23.46 (34.10)	17.66 (32.66)	28.32 (34.65)	.87
WIAT CIWS SS	100.35 (17.13)	97.96 (16.55)	103.36 (17.44)	NA	NA	NA	NA
Narrative CIWS	33.99 (36.83)	28.31 (31.25)	41.33 (42.00)	17.48 (31.69)	13.67 (31.88)	20.64 (31.32)	.85
Pet CIWS	37.31 (36.71)	29.40 (31.85)	46.86 (39.94)	26.58 (34.84)	18.74 (29.86)	33.30 (37.43)	.79
Writing productivity indicators							
WIAT no. of words	82.38 (33.66)	75.83 (31.17)	90.58 (34.97)	74.77 (34.99)	66.36 (34.90)	81.88 (33.59)	.87
WIAT no. of words SS	109.00 (13.49)	106.51 (13.14)	112.15 (13.33)	NA	NA	NA	NA
Narrative no. of words	89.44 (38.38)	82.01 (34.08)	99.03 (41.52)	82.36 (38.12)	73.79 (36.56)	89.47 (38.06)	.84
Pet no. of words	80.05 (36.92)	72.52 (35.24)	89.14 (37.01)	73.86 (41.14)	64.33 (39.14)	82.11 (41.21)	.76
WIAT no. of ideas	12.07 (5.04)	11.28 (4.80)	13.06 (5.18)	11.73 (5.51)	10.44 (5.34)	12.82 (5.43)	.80
Narrative no. of ideas	15.62 (6.69)	14.30 (6.02)	17.31 (7.14)	14.26 (6.80)	12.76 (6.46)	15.50 (6.84)	.78
Pet no. of ideas	12.95 (5.84)	11.74 (5.52)	14.41 (5.91)	11.69 (6.01)	10.47 (5.84)	12.76 (5.98)	.69

Note. WIAT = Wechsler Individual Achievement Test (3rd edition); SS = standard score; WJ-III = Woodcock-Johnson Tests of Achievement (3rd edition); Narrative = Test of Narrative Language; Pet = pet prompt; CBM = curriculum-based measurement; CWS = correct word sequences; IWS = incorrect word sequences; CIWS = correct minus incorrect word sequences. Loadings were for the following latent variables: writing quality, (curriculum-based measurement) CBM writing, and productivity; NA = not applicable. The loadings of the WIAT theme and organization and WJ-III Writing Fluency raw were those when they were considered as indicators of the writing quality.

^a Words written + theme and organization.

not presented for this grade. In the WIAT writing composition, the standard score is a composite of the standard scores from the number of words written and the theme development and organization standard scores. The standard score in the WIAT writing task is in the average range albeit slightly in the high average for children in Grade 3 (mean standard score [SS] = 107.92, $SD = 13.74$). Standard scores in the WJ-III Writing Fluency task was in the average range as well (mean SS = 98.28 and 95.09 for Grades 3 and 2, respectively). The WIAT Grammar Score, which is CIWS CBM writing, was in the average range (mean SS = 100.25) for students in Grade 3. However, note that the standard scores for the Grammar Score should be viewed with caution due to slight differences in scoring CIWS between WIAT and our approach following previous studies (e.g., McMaster et al., 2009).

Table 2 displays descriptive statistics for language and literacy predictors by grade and gender. In the language measures, mean performance was in the average range—from 8.65 on the Test of Narrative Language Narrative Comprehension subtest to 99.13 on the WJ-III Picture Vocabulary task. Children's reading skills, spelling, and alphabet writing fluency were also in the average range. Correlations are presented in Table 3 for writing variables and in Table 4 for language and cognitive variables. Preliminary analysis showed that the patterns of relations were highly similar for children in Grades 2 and 3, and thus, results from combined data are presented. The writing quality variables tended to be

moderately and statistically significantly related to each other while writing productivity variables (number of words and number of ideas) were highly related to each other. Given that RAN has not been examined in relation to writing in previous studies with English-speaking children, correlations of RAN to writing scores are presented in Table 3. RAN was weakly to moderately related to all the writing variables ($-.24 \leq rs \leq -.43$). Language and cognitive variables in Table 4 were all statistically significantly correlated in expected directions.

Dimensionality of Writing

In order to examine the dimensionality captured in the various writing evaluation measures, we conducted a series of analysis. First, we confirmed the hypothesized factor structure using CFA models (measurement models) for the writing quality and productivity. Writing quality and productivity were deemed to be a good place to start because previous studies indicated that they are dissociable dimensions and their indicators are fairly well understood (Kim, Al Otaiba, et al., 2014; Kim, Park, & Park, 2013; Puranik et al., 2008; Wagner et al., 2011). Second, we examined measurement model (i.e., CFA) of the CBM writing scores and its relation to writing quality and writing productivity dimensions. Finally, we examined whether the WJ-III Writing Fluency is best

Table 2
Means (Standard Deviations) of Language and Literacy Predictors by Gender

Variable	Grade 3			Grade 2			Loadings
	Entire sample	Males	Females	Entire sample	Males	Females	
OWLS raw	87.09 (10.71)	87.29 (10.20)	86.84 (11.37)	82.54 (11.00)	81.15 (11.25)	83.74 (10.69)	.74
OWLS SS	98.09 (13.48)	98.12 (13.01)	98.05 (14.12)	101.40 (12.47)	99.57 (12.59)	102.98 (12.19)	NA
TNL Narrative Comprehension raw	28.34 (4.60)	27.91 (4.67)	28.87 (4.47)	26.60 (4.89)	25.79 (5.29)	27.29 (4.42)	.70
TNL Narrative Comprehension SS	8.65 (3.06)	8.33 (2.92)	9.04 (3.21)	8.33 (2.70)	7.86 (2.70)	8.73 (3.21)	NA
WJ-III Picture Vocabulary raw	23.24 (3.16)	23.22 (3.32)	23.28 (2.96)	21.50 (3.19)	21.58 (3.07)	21.43 (3.30)	.75
WJ-III Picture Vocabulary SS	99.13 (10.41)	99.00 (10.89)	99.30 (9.79)	98.74 (10.61)	98.87 (9.92)	98.64 (11.21)	NA
WJ-III LWID raw	50.26 (6.64)	50.04 (6.74)	50.54 (6.53)	44.37 (7.31)	43.84 (7.47)	44.82 (7.18)	.85
WJ-III LWID SS	104.84 (11.04)	104.40 (11.21)	105.41 (10.85)	105.89 (11.41)	104.84 (11.34)	106.80 (11.44)	NA
Sight Word Efficiency raw	62.88 (11.59)	61.79 (10.77)	64.20 (12.45)	54.45 (13.19)	53.43 (14.24)	55.31 (12.22)	.89
Sight Word Efficiency SS	96.26 (15.02)	94.50 (13.95)	98.43 (16.04)	99.57 (15.20)	98.05 (16.29)	100.84 (14.16)	NA
WIAT-ORF1	60.09 (23.48)	61.70 (23.14)	58.04 (23.86)	65.67 (28.27)	67.04 (29.86)	64.50 (26.89)	NA
WIAT-ORF2	71.69 (28.42)	74.06 (27.75)	68.70 (29.91)	77.83 (41.70)	83.45 (49.13)	73.02 (33.54)	NA
WIAT-ORF weighted raw	105.17 (35.81)	101.75 (34.41)	109.46 (37.21)	88.57 (33.43)	85.10 (33.81)	91.44 (32.96)	.88
WIAT-ORF SS	103.49 (14.98)	101.85 (14.60)	105.54 (15.25)	99.23 (13.58)	97.37 (13.92)	100.78 (13.14)	NA
TOSREC raw	25.24 (9.27)	24.21 (9.39)	26.48 (9.02)	26.09 (9.75)	24.80 (9.90)	27.22 (9.52)	.79
TOSREC SS	101.22 (16.37)	99.36 (16.61)	103.48 (15.85)	98.64 (15.22)	96.64 (15.33)	100.37 (14.97)	NA
WJ-III Passage Comprehension raw	25.77 (3.49)	25.66 (3.52)	25.91 (3.47)	23.44 (3.92)	23.07 (3.95)	23.76 (3.87)	.80
WJ-III Passage Comprehension SS	95.38 (9.48)	95.05 (9.56)	95.82 (9.41)	97.51 (9.45)	96.44 (9.40)	98.45 (9.42)	NA
WIAT Alphabet Writing Fluency raw	17.57 (6.35)	17.32 (6.40)	17.87 (6.31)	15.91 (6.04)	15.76 (6.10)	16.03 (6.00)	NA
WIAT Alphabet Writing Fluency SS	104.74 (18.95)	104.12 (18.33)	105.53 (19.77)	104.62 (17.31)	104.22 (16.85)	104.97 (17.77)	NA
WJ-III Spelling raw	32.63 (5.88)	32.57 (5.88)	32.70 (5.90)	28.85 (5.51)	28.56 (5.88)	29.09 (5.20)	NA
WJ-III Spelling SS	102.75 (14.45)	102.50 (14.70)	103.06 (14.20)	102.58 (13.92)	101.46 (14.40)	103.53 (13.50)	NA
Story copying: letters correct	36.21 (15.96)	33.51 (13.99)	39.60 (17.62)	27.55 (11.59)	26.17 (12.35)	28.75 (10.80)	NA
SWAN Attention	34.36 (10.13)	33.01 (9.73)	36.08 (10.42)	36.68 (11.90)	33.19 (11.77)	39.62 (11.24)	NA
RAN Time	28.40 (7.15)	28.45 (6.76)	28.35 (7.64)	32.53 (8.88)	33.12 (9.82)	32.03 (8.01)	NA
RAN SS	99.88 (12.63)	99.37 (11.98)	100.54 (13.43)	99.47 (12.84)	98.57 (13.24)	100.23 (12.49)	NA

Note. Raw = raw score; OWLS = Oral and Written Language Scale; TNL = Test of Narrative Language; SS = standard score; WJ-III = Woodcock-Johnson Tests of Achievement (3rd edition); LWID = Letter Word Identification; WIAT = Wechsler Individual Achievement Test (3rd edition); ORF = Oral Reading Fluency; TOSREC = Test of Silent Reading Efficiency and Comprehension; SWAN = Strengths and Weaknesses of ADHD Symptoms and Normal Behavior Rating Scale; RAN = Rapid Automatized Naming Test. Loadings were for oral language latent variable (OWLS, TNL, & WJ Picture Vocabulary) and reading latent variable (WJ Letter Word Identification, Test of Word Reading Efficiency Sight Word Efficiency, WIAT ORF, TOSREC, and WJ Passage Comprehension); NA = not applicable.

Table 3
Correlations Among Writing Variables

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1. WIAT TDTO	1.00																			
2. WIAT Q Ideas	.48	1.00																		
3. Narrative Q Ideas	.45	.48	1.00																	
4. Pet Q Ideas	.38	.44	.34	1.00																
5. WIAT Q Org	.57	.42	.40	.33	1.00															
6. Narrative Q Org	.43	.37	.66	.29	.43	1.00														
7. Pet Q Org	.38	.35	.40	.46	.43	.43	1.00													
8. WJ-III Writing Fluency	.46	.45	.42	.28	.48	.47	.39	1.00												
9. WIAT CIWS	.55	.52	.42	.40	.47	.43	.42	.53	1.00											
10. Narrative CIWS	.46	.42	.48	.37	.40	.49	.45	.48	.73	1.00										
11. Pet CIWS	.45	.42	.41	.53	.44	.42	.44	.44	.73	.69	1.00									
12. WIAT no. of words	.47	.66	.40	.43	.26	.23	.23	.35	.50	.30	.39	1.00								
13. Narrative no. of words	.44	.55	.58	.38	.22	.36	.27	.39	.42	.41	.40	.72	1.00							
14. Pet no. of words	.36	.44	.31	.63	.20	.19	.27	.23	.33	.25	.50	.66	.63	1.00						
15. WIAT no. of ideas	.37	.58	.36	.32	.17	.18	.15	.27	.38	.22	.28	.89	.68	.58	1.00					
16. Narrative no. of ideas	.42	.54	.57	.34	.22	.37	.27	.38	.41	.39	.38	.68	.94	.57	.65	1.00				
17. Pet no. of ideas	.35	.43	.35	.62	.23	.32	.26	.34	.27	.50	.59	.58	.92	.54	.55	.100				
18. WIAT %CWS	.44	.29	.33	.28	.41	.37	.42	.46	.85	.66	.61	.20	.22	.15	.14	.22	.19	.100		
19. Narrative %CWS	.35	.19	.25	.21	.33	.32	.35	.59	.83	.55	.08 ^{ns}	.08 ^{ns}	.06 ^{ns}	.03 ^{ns}	.09 ^{ns}	.09 ^{ns}	.71	.1.00		
20. Pet %CWS	.35	.26	.35	.22	.43	.38	.43	.62	.61	.78	.14	.20	.09	.07 ^{ns}	.21	.14	.69	.65	1.00	
21. RAN	-.36	-.37	-.40	-.29	-.30	-.32	-.29	-.43	-.40	-.36	-.37	-.41	-.43	-.34	-.31	-.42	-.33	-.35	-.24	-.35

Note. All coefficients are statistically significant at .05 level except those with superscript "ns." WIAT = Wechsler Individual Achievement Test (3rd edition); TDTO = theme development and text organization; Q = quality; Org = organization; WJ-III = Woodcock-Johnson Test of Achievement (3rd edition); Narrative = narrative prompt; Pet = pet prompt; CIWS = correct minus incorrect word sequences; CWS = correct word sequences; RAN = Rapid Automatized Naming Test.

Table 4
Correlations Among Language and Cognitive Variables

Variable	1	2	3	4	5	6	7	8	9	10	11	12
1. OWLS	1.00											
2. TNL Narrative Comprehension	.51	1.00										
3. WJ-III Picture Vocabulary	.55	.51	1.00									
4. WJ-III Letter Word Identification	.45	.40	.54	1.00								
5. TOWRE Sight Word Efficiency	.35	.34	.43	.76	1.00							
6. WIAT ORF weighted	.40	.38	.47	.72	.80	1.00						
7. TOSREC	.42	.43	.48	.66	.69	.72	1.00					
8. WJ-III Passage Comprehension	.52	.46	.63	.74	.69	.67	.67	1.00				
9. WIAT Alphabet Writing Fluency	.21	.25	.24	.38	.39	.34	.35	.31	1.00			
10. WJ-III Spelling	.36	.28	.42	.78	.68	.66	.57	.59	.39	1.00		
11. Story copying: letters correct	.26	.25	.23	.35	.39	.40	.32	.36	.43	.41	1.00	
12. SWAN Attention	.36	.37	.32	.43	.44	.52	.59	.44	.26	.47	-.27	1.00
13. RAN	-.16	-.17	-.18	-.48	-.69	-.52	-.41	-.43	-.35	-.46	-.36	-.27

Note. All coefficients are statistically significant at .001 level. OWLS = Oral and Written Language Scale; TNL = Test of Narrative Language; WJ-III = Woodcock-Johnson Tests of Achievement (3rd edition); TOWRE = Test of Word Reading Efficiency. WIAT = Wechsler Individual Achievement Test (3rd edition); ORF = Oral Reading Fluency; TOSREC = Test of Silent Reading Efficiency and Comprehension; SWAN = Strengths and Weaknesses of ADHD Symptoms and Normal Behavior Rating Scale; RAN = Rapid Automatized Naming Test.

described as an indicator of the writing quality, productivity, CBM writing, or as a separate observed variable.

We hypothesized that the theme and organization score of the WIAT composition task would capture the writing quality along with the idea development and organization aspects of the adapted 6 + 1 Trait Rubric because the theme and organization of the WIAT task evaluates idea development and structural aspects of written composition. CFA confirmed the hypothesis: The model fit was good: $\chi^2(13) = 72.92$, $p < .001$; CFI = .95; TLI = .92; RMSEA = .097; and SRMR = .038. Factor loadings are presented in Table 1. Based on preliminary analysis, error covariance was allowed between the theme and organization and the 6 + 1 organization score. The CFA model for writing productivity using number of words written and number of ideas yielded an excellent fit: $\chi^2(6) = 34.18$, $p < .001$; CFI = .99; TLI = .98; RMSEA = .10; and SRMR = .01.

To examine the dimensionality of the variables derived from the CBM scoring approaches, we fit two CFA models (two latent variables in which the CIWS variable is dissociable from %CWS vs. one latent variable in which both CIWS and %CWS capture a single latent variable), and we compared model fits. The model fit for a single dimension was slightly better, $\Delta\chi^2 = 5.87$, $\Delta df = 1$, $p = .02$. However, the CIWS and %CWS were very highly correlated when modeled separately ($r = .97$). Therefore, it appeared reasonable to model both the CIWS and %CWS as a single CBM latent variable (noted as CBM writing scoring hereafter) in

subsequent analysis. Table 5 shows comparison of CFA model fits for alternative models examining whether writing quality, productivity, and CBM writing were best considered as three dissociable variables or two dissociable variables, or as a single variable. Results showed the three-latent-variable model describes the data best compared with the other alternative models, $\Delta\chi^2 \geq 201.21$, $ps < .001$.

Next, we examined whether the WJ-III Writing Fluency is best described as an indicator of the identified dimensions of writing (writing quality, productivity, CBM) or is better described as a separable variable. When we fit a model in which the WJ-III Writing Fluency task was considered as a separate variable from the other three (i.e., writing quality, productivity, and CBM writing), the fit was acceptable: $\chi^2(151) = 1055.14$, $p < .001$; CFI = .90; TLI = .88; RMSEA = .011; SRMR = .08. The WJ-III Writing Fluency task correlated most strongly with the writing quality at .67, followed by .59 with CBM writing, and .46 with productivity. When the WJ-III Writing Fluency task was considered as an indicator of CBM writing or productivity, the model fits were statistically significantly worse ($ps < .001$). When a CFA model was fit in which the WJ-III Writing Fluency was considered as an indicator of writing quality, the model fit was not different from the separate dimension model, that is, the four-factor model; $\Delta\chi^2$, $\Delta df = 2 = 5.82$, $p = .054$. Therefore, based on these results and for parsimony, the WJ-III Writing Fluency task is considered as an indicator of writing quality.

Table 5
Model Fit Indices for Alternative Models

Model No. and description	$\chi^2 (df)$	CFI	TLI	RMSEA	SRMR	Comparison to Model 1: $\Delta\chi^2$, Δdf (p)
1. Three latent variables (quality, productivity, CBM)	1061.00 (153)	.90	.88	.11	.083	
2. Two latent variables (quality + CBM, productivity)	1262.21 (155)	.88	.85	.12	.098	201.21, 1 ($p < .001$)
3. Two latent variables (productivity + CBM, quality)	1702.65 (155)	.83	.79	.14	.12	641.65, 1 ($p < .001$)
4. Two latent variables (quality + productivity, CBM)	1344.67 (155)	.87	.84	.13	.106	283.67, 1 ($p < .001$)
5. One latent variable (quality + productivity + CBM)	1727.22 (156)	.83	.79	.14	.121	666.22, 2 ($p < .001$)

Note. CFI = comparative fit index; TLI = Tucker-Lewis index; RMSEA = root-mean-square error of approximation; SRMR = standardized root-mean-square residuals; CBM = curriculum-based measurement.

In summary, CFA analysis revealed the following three dimensions for the writing outcomes: writing quality, writing productivity, and CBM writing. The writing quality dimension was strongly related to CBM writing at .82 and to writing productivity at .75. Writing productivity and CBM writing were moderately correlated at .54.

Language and Cognitive Predictors of Writing Quality, Writing Productivity, and CBM Writing

Factor scores of the three writing dimensions (writing quality, productivity, and CBM writing) from CFA results were extracted from Mplus ($SDs = 1.83, 25.74$, and 28.74 , for writing quality, productivity, and CBM, respectively; means are 0), and these three dimensions were used in subsequent multilevel modeling with SAS 9.3. In addition, latent variables were created for the predictors with multiple measures (i.e., oral language and reading) using CFA. Factor loadings were high (see Table 2), and model fits were excellent (not shown). Then, factor scores of these language and reading latent variables were used in the multilevel models.

First, unconditional models without any predictors were fit for the three writing outcomes to parse out amount of variance attributable to individuals, classrooms, and schools. Intraclass correlations were as follows: (a) writing quality, .16 at the school level and .05 at the classroom level; (b) writing productivity, .07 at the school level, but 0 at the classroom level; and (c) CBM writing, .16 at the school level and .15 at the classroom level. In other words, approximately 16% of the total variance in writing quality, 7% in writing productivity, and 16% in CBM writing were due to school differences, whereas approximately 5% of the total variance in writing quality, 0% in writing productivity, and 15% were due to

differences among classrooms. In the subsequent analysis, a three-level model (school, classroom, and individual) was carried out for the writing quality and CBM outcomes, whereas a two-level model (school and individual) was constructed for the writing productivity outcome because of lack of variance at the classroom level in writing productivity.

We then fit models (M1) to examine unique correlates of writing quality, writing productivity, and CBM writing (Research Question 2). As shown in Tables 6 and 7, for writing quality, all the language and cognitive predictors were statistically significant after accounting for children's age: children's oral language ($p = .004$), reading ($p < .001$), spelling ($p < .001$), letter writing automaticity ($p = .048$), story copying ($p < .001$), RAN ($p = .005$), and attention ($p = .03$). After accounting for all these variables, no variance remained at the classroom and school levels. For the writing productivity, individual differences in reading ($p = .002$) and timed tasks such as letter writing automaticity ($p = .004$), story copying ($p < .001$), and RAN ($p < .001$) were related, whereas oral language, spelling, and attention were not ($ps \geq .24$). Finally, for the CBM writing scoring outcome, children's reading ($p < .001$), spelling ($p < .001$), story copying ($p < .001$), and attention ($p = .02$) remained statistically significant, whereas oral language, letter writing automaticity, and rapid automatized naming did not ($ps \geq .43$).

Gender and Writing

To address the third research question of gender gap, first, we included children's gender as the main predictor in addition to the age control variable for each writing outcome. This allowed us to see whether gender differences were found after account-

Table 6

Results of Multilevel Models: Writing Quality and Writing Productivity Predicted by Students' Language and Literacy Skills, Attention, and Gender

Variable	Writing quality			Writing productivity		
	M1	M2	M3	M1	M2	M3
Fixed effects						
Intercept	-2.35 (0.83)***	0.34 (1.14)	-2.24 (0.81)*	19.42 (16.28)	-3.11 (15.40)	21.71 (15.88)
Age in months	-0.04 (0.08)	-0.02 (0.14)	-0.02 (0.08)	-2.52 (1.59)	0.96 (1.86)	-2.06 (1.55)
Male	NA	-0.71 (0.15)***	-0.41 (0.10)***	NA	-11.80 (2.22)***	-8.70 (1.93)***
Reading	0.10 (0.02)***		0.10 (0.02)***	0.95 (0.002)**		0.91 (0.30)**
Oral language	0.03 (0.01)**		0.03 (0.009)**	-0.15 (0.18)		-0.11 (0.17)
WJ-III spelling	0.05 (0.01)***		0.06 (0.01)***	0.52 (0.18)		-0.13 (0.24)
WIAT letter writing	0.02 (0.01)*		0.02 (0.009)*	0.52 (0.18)**		0.55 (0.17)**
Story copying	0.03 (0.004)***		0.03 (0.004)***	0.54 (0.08)***		0.51 (0.08)***
SWAN attention	0.01 (0.005)*		0.008 (0.006)	0.13 (0.11)		-0.03 (0.11)
RAN	-0.02 (0.008)**		-0.02 (0.008)**	-0.73 (0.15)***		-0.75 (0.15)***
Variance components						
School	0	0.57	0	17.42	43.16	15.65
Classroom	0	0.25	0	NA	NA	NA
Children	1.07	2.42	1.03	366.44	580.09	349.77
-2LL	12.06.2	1929.9	1190.8	3641.8	4568.1	3621.8
AIC	1226.2	1941.7	1212.8	3663.8	4578.1	3645.8

Note. Standard deviations are in parentheses. M1 = Model 1 (examines the relation of language and cognitive skills); M2 = Model 2 (examines the relation of gender and writing); M3 = Model 3 (examines the relation of gender and writing after accounting for language and cognitive skills); NA = not applicable; WJ-III = Woodcock-Johnson Tests of Achievement (3rd edition); WIAT = Wechsler Individual Achievement Test; SWAN = Strengths and Weaknesses of ADHD Symptoms and Normal Behavior Rating Scale; RAN = Rapid Automatized Naming Test; -2LL = log-likelihood; AIC = Akaike information criterion.

* $p < .05$. ** $p < .01$. *** $p < .001$.

Table 7

Results of Multilevel Models: Curriculum-Based Measurement (CBM) Writing Scoring Predicted by Students' Language and Literacy Skills, Attention, and Gender

	CBM writing		
	M1	M2	M3
Fixed effects			
Intercept	-68.45 (13.91)***	14.35 (17.89)	-66.39 (13.74)***
Age in months	-0.67 (1.33)	-1.40 (2.15)	-0.35 (1.32)
Male	NA	-10.67 (2.36)***	-6.35 (1.70)***
Reading	1.48 (0.27)***		1.45 (0.26)***
Oral language	0.12 (0.16)		0.16 (0.15)
WJ-III spelling	-0.04 (0.16)		1.88 (0.21)***
WIAT letter writing	1.81 (0.22)***		-0.02 (0.15)
Story copying	0.26 (0.07)***		0.24 (0.07)***
SWAN attention	0.30 (0.09)***		0.23 (0.10)*
RAN	-0.03 (0.13)		-0.04 (0.13)
Variance components			
School	5.12	150.82	6.09
Classroom	0	129.65	0
Children	284.10	528.60	274.05
-2LL	3528.5	4648.2	3514.8
AIC	3550.5	4660.2	3539.6

Note. M1 = Model 1 (examines the relation of language and cognitive skills); M2 = Model 2 (examines the relation of gender and writing); M3 = Model 3 (examines the relation of gender and writing after accounting for language and cognitive skills); NA = not applicable; WJ-III = Woodcock-Johnson Tests of Achievement (3rd edition); WIAT = Wechsler Individual Achievement Test; RAN = Rapid Automatized Naming Test; -2LL = log-likelihood; AIC = Akaike information criterion.

* $p < .05$. *** $p < .001$.

ing for age, and if so, how large the gaps were before including any potential explanatory variables. As shown in the second models in Tables 6 and 7, in all the writing outcomes, boys had statistically significantly lower scores after accounting for age. In writing quality, boys scored, on average, 0.39 standard deviation lower than girls. In writing productivity, boys' score was, on average, lower than girls by 0.46 standard deviation, and in CBM writing, boys' score was 0.37 standard deviations lower than girls.

Language and cognitive variables were then included in the models to investigate whether gender differences in writing score persisted or disappeared after controlling for these language and cognitive variables. Results in Tables 6 and 7 (M3) show that boys continued to have lower mean scores in writing even after accounting for all the included language and cognitive variables. However, the effect sizes were reduced by approximately quarter to a third compared with those in the initial models: the effect sizes were .22 in writing quality, .34 in writing productivity, and .22 in CBM writing. In other words, the included language and literacy predictors explained the gender gap in writing outcomes to some extent, but the relation between gender and writing was not completely mediated by the included language and cognitive skills. It is of note that the relation of language and cognitive skills to the three writing outcomes essentially remained the same between M1 (before controlling for gender) and M3 (after accounting for gender). However, an exception was found for attention, which was no longer related to writing quality once gender was taken into consideration in addition to language and cognitive skills and age.

Discussion

In the present study, we investigated the dimensionality of writing, predictors of writing, and gender differences, using a large data set from second and third grade students in the United States. Findings showed that writing quality, writing productivity, and CBM writing (CIWS and %CWS) were dissociable dimensions, at least for children in Grades 2 and 3. Furthermore, unique predictors of each dimension differed.

In conjunction with previous studies (Kim, Al Otaiba, et al., 2014; Puranik et al., 2008; Wagner et al., 2011), the present findings suggest that writing is not a single dimension but is composed of multiple dimensions. Theoretically, the writing quality and productivity dimensions describe skills that are hypothesized to be products of two key components in writing, namely, ideation and transcription skills (Juel et al., 1986). Idea development and organization aspects, theme and organization scores in WIAT, and the WJ-III Writing Fluency task all captured the writing quality dimension, whereas number of words written and number of ideas captured the writing productivity dimension. These findings confirm previous studies about the dissociability of writing quality and productivity (Kim, Al Otaiba, et al., 2014; see also Puranik et al., 2008, and Wagner et al., 2011), but extend our understanding by demonstrating that the theme and organization score of WIAT and the sentence level WJ-III Writing Fluency tasks capture writing quality. It is interesting that the WJ-III Writing Fluency task was more strongly related to writing quality than to writing productivity or CBM writing and was best described as an indicator of writing quality. This result suggests that the accuracy and rate at which children can construct sentences is

likely to be an indicator of writing quality but not writing productivity or CBM writing, at least at this stage of writing development. It might be that the WJ-III Writing Fluency task captures efficiency of children's transcription skills and sentence production skills (an oral language skill), both of which are important for written composition. It is plausible that this efficiency enables children to focus on higher order processes such as idea expression and organization.

Although CBM writing measures and coding methods have been examined for reliability and validity (see Graham et al., 2011; McMaster & Espin, 2007, for a review), the nature of their theoretical construct and dimensionality has been nebulous. In the present study, CBM writing scores captured a dissociable dimension from writing quality and productivity and was strongly associated with the quality of writing at .82 and moderately associated with writing productivity at .54. It should be noted that in the present study, we included two scoring tools that are unique to CBM, CIWS and %CWS, although CBM writing scores also include other indicators such as total number of words written. This latter variable was conceptualized as a productivity indicator in the present study according to previous studies and findings (e.g., Abbott & Berninger, 1993; Graham et al., 1997; Kim, Al Otaiba, et al., 2014; Wagner et al., 2011). Whether the separate CBM writing dimension in the present study should be conceptualized as a global outcome measure of children's writing skill or writing fluency, or as another construct is beyond the scope of the present study. As noted earlier, CBM writing was recently theorized as writing fluency, which was defined as the ease to generate written text. According to automaticity and information processing theories (e.g., LaBerge & Samuels, 1974; Posner & Snyder, 1975), fluency (or automaticity) is required so that cognitive resources such as attention and working memory can be used for higher order cognitive resources. Applying this to writing development, efficiency in generating ideas and transcribing those ideas into written texts would allow a writer to focus on aspects such as presenting ideas in an organized, clear, and rich manner to enhance writing quality. The two CBM writing variables used in the present study (CIWS and %CWS) appear to operationalize writing fluency well because both capture not just the amount of writing but efficiency (accuracy and amount). In addition, CIWS and %CWS tend to have highest validity evidence (e.g., Amato & Watkins, 2011; McMaster & Espin, 2007). One way to validate CBM writing measures (at least CIWS and %CWS) as indicators of writing fluency is to examine how data fit this theoretical hypothesis. Specifically, Ritchey et al. (in press) hypothesized that writing fluency includes text generation and transcription, which is aligned well with the simple view of writing (Juel et al., 1986) and the not-so-simple view of writing (Berninger & Winn, 2006). Therefore, text generation and transcription skills are component skills of writing fluency (i.e., CBM writing), which then would predict the criterion measure of writing such as writing quality. In other words, the CBM writing measures should mediate, at least partially, the relations of the text generation and transcription to the criterion measure of writing. Effort is under way to investigate this hypothesis by the current research team.

Another piece of evidence about multiplicity of writing dimensions comes from differential relations of language and cognitive skills to the three dimensions. In the model after accounting for gender (Models 3), whereas reading, letter writing fluency, and

rapid automatized naming were related to both writing quality and productivity, oral language and spelling were related only to writing quality, but not to writing productivity. In addition, attention was related to the CBM writing outcome over and above the other variables in the model. Interestingly, although CIWS and %CWS do take into consideration grammatical accuracy, oral language skill did not uniquely influence the CBM writing. It is notable that reading was a consistent predictor for all three dimensions, underscoring the importance of early reading skill in early writing, even after accounting for other variables in the model. These results add to the increasing evidence of the relation between reading and writing, particularly in the elementary years (Berninger et al., 2002; Kim, Al Otaiba, et al., 2013, in press; Shanahan, 2006; Shanahan & Lomax, 1986). Reading has been hypothesized to play a role during the process of self-monitoring during planning and revision as children have to assess and plan for revisions (Hayes, 1996; McCutchen, Francis, & Kerr, 1997). Additionally, reading skills might contribute to the quality of writing by way of reading experiences—better readers read more, and greater amount of reading might help children with idea generation from increased background knowledge and better organization of ideas (Berninger et al., 2006).

Transcription skills also tended to be consistently related to the writing outcomes. Spelling skill was related to writing quality and CBM writing, and letter writing automaticity was related to writing quality and writing productivity. These findings confirm previous studies about the role of transcription skills in writing, as they are needed not only to encode ideas into written language but also to allow cognitive resources to be used for higher order writing processes (Abbott & Berninger, 1993; Berninger et al., 1997; Graham et al., 1997). It is noteworthy that compared with the letter writing task, the story copying task was related to all three of the writing outcomes after accounting for the other variables in the model, suggesting that story copying captures processes beyond those captured by the alphabet letter writing task. Story copying may involve a greater extent of processing capacity (e.g., working memory) to hold and process words and sentences as it is a discourse-level text, whereas a letter writing task is simply retrieval of letters from memory. Future studies are needed to replicate the results and any potential sources of differences between letter writing and story copying tasks.

Attention was another cognitive skill that was hypothesized to be important for writing (Berninger & Winn, 2006), and it was related to writing quality and CBM writing in the present study, confirming previous findings for children in first grade and in kindergarten (Kent et al., in press; Kim, Al Otaiba, et al., 2013). Interestingly, once children's gender was accounted for, attentiveness was not related to writing quality although its relation remained for the CBM writing outcome. These results suggest that gender may mediate the relation between attention and writing quality. Previous studies did not include gender as a covariate in examining the role of attention in writing. Future studies are needed to investigate the precise role of attention in writing development including reasons why attention matters for CBM writing. This is important for typically developing students but also for students with ADHD, as boys are more commonly diagnosed than girls (Arcia & Conners, 1998; Levy, Hay, Bennett, & McStephen, 2005).

RAN was weakly to moderately related to various writing scores in bivariate correlations. Once other language and literacy skills were accounted for, RAN was independently related to writing quality and productivity, and to our knowledge, this was the first study to examine this relation in English. On the one hand, our findings converge with two previous studies in another orthography, Chinese (Chan et al., 2006; Ding et al., 2010). On the other hand, however, they are discrepant from those of a third study with Chinese children in which RAN was not related to writing once transcription skills were accounted for (Yan et al., 2012). If RAN captures mostly automaticity of letter retrieval, then its influence should be largely shared with handwriting automaticity tasks such as letter writing automaticity and story copying. Given its relations to writing quality and productivity, it appears that RAN captures processes beyond handwriting fluency. According to the multi-component account of RAN (Wolf & Bowers, 1999; Wolf & Denkla, 2005), RAN includes processes for visual, orthographic, and verbal processing, and this integration process might be a factor that drives the independent relation of RAN to writing quality and productivity over and above the other language and literacy skills.

These results of multiple dimensions and associated predictors offer important implications for instruction and assessment practices. Instructionally, teachers may target different aspects and skills to ensure student progress on all areas of writing. In addition, if data suggest that a child has weaknesses that may impact a particular writing dimension, teachers may target skills in that area of interest. For instance, if the teacher is mostly interested in improving children's writing quality, the teacher may want to model and introduce strategies to help students focus on the development and organization of ideas and expressing generated ideas with appropriate language. Also it is worthy to note that to improve writing quality, instructional attention is needed in multiple aspects such as oral language, reading, transcription skills, RAN, and attention, given that the quality of writing was predicted by the wide array of language and literacy and cognitive assessments. If the teacher is particularly concerned about the children's productivity, the teacher may focus more on transcription-related skills, given their roles in writing productivity. The teacher could target spelling, sentence writing fluency, or other related transcription skills.

Furthermore, if the teacher's primary goal is progress monitoring in writing, the CBM writing scores appear most appropriate for two reasons. First, although CBM and writing quality appear to be separable dimensions, CBM writing scores give a general idea about writing quality, given a strong relation between writing quality and CBM writing scores ($r = .82$). Second, CBM writing scores have been shown to be reliable and sensitive to growth captured within a relatively short span of time, which is important due to frequent assessments (e.g., 2 weeks; Graham et al., 2011; Lembke et al., 2003; McMaster & Espin, 2007; McMaster et al., 2009, 2011). In contrast, writing quality may be less appropriate for frequent assessments because writing quality indicators, which are typically evaluated on a rating scale, are not likely to be as sensitive as CBM writing measures in capturing changes during a short period. This speculation, however, requires a future study.

Finally, confirming previous studies (Berninger & Fuller, 1992; Knudson, 1995; National Center for Education Statistics, 2011), boys in the present study performed more poorly in all the three

writing dimensions, with effect sizes ranging from .37 to .46. Results further showed that gender differences were explained by the included language and cognitive skills as the effect sizes in gender differences were reduced by approximately quarter to a third when these variables were taken into account. In other words, the language and cognitive variables included in the present study partially explained writing performance differences between boys and girls. On the other hand, these results indicate that gender differences persisted in all of the three writing outcomes even after accounting for these language and cognitive skills. These findings indicate that studies are needed to expand the understanding of potential causes of gender gaps in writing. Given findings of a previous study that even in Grade 1, boys engage in less writing (Graham et al., 2007), it would be informative to investigate how attitude toward writing together with language and literacy variables explains gender gaps in writing, and whether attitude is malleable. Additionally, other potential sources of gender gaps (e.g., persistence in writing; McKenna, Kear, & Ellsworth, 1995) need to be investigated in future studies.

Limitations and Conclusion

One of the limitations of the present study is that many children in the present study came from low-income family backgrounds from one midsized city in the Southeast. In addition, the children were primarily African Americans and Whites, with virtually no English-language learners. Although their writing performance was in the average range in standardized and normed writing assessments, future research is needed to determine whether similar results are found for children from different socioeconomic and linguistic backgrounds. Further understanding is also required regarding the CBM writing scoring dimension. Many studies have shown technical adequacy and the utility of CBM in screening and progress monitoring of elementary grade children's writing. Recent efforts in theoretical conceptualization (e.g., McMaster & Espin, 2007; Ritchey et al., in press) are in the right direction to help the field gain better understanding of this dimension of writing. Finally, there are other types of evaluative approaches to written compositions and predictors of writing skills that were not included in the present study. For instance, text elements (e.g., presence of text structural elements such as topic sentence and supporting details; Kulikowich, Mason & Brown, 2008; Wagner et al., 2011) and spelling and writing conventions (e.g., punctuation and handwriting) were not examined in the present study. In addition, motivational, discourse knowledge, and cognitive factors (e.g., strategic writing) have been shown to be related to writing skills (e.g., Bruning & Horn, 2000; Graham et al., 2005; Hidi & Boscolo, 2006; Limpo & Alves, 2013; Pajares, 2003; Olinghouse & Graham, 2009) but were not examined in the present study.

Overall, the findings of the present study suggest that writing quality, writing productivity, and CBM writing (composed of CIWS and %CWS) are separate dimensions for children in Grades 2 and 3 and that the relations of language and literacy variables differed for various writing outcomes. In addition, gender differences persist even after accounting for language and cognitive skills. Future research is needed to replicate the present study and to further expand researchers' understanding about skills that influence children's writing development.

References

Abbott, R. D., & Berninger, V. W. (1993). Structural equation modeling of relationships among developmental skills and writing skills in primary- and intermediate-grade writers. *Journal of Educational Psychology, 85*, 478–508. doi:10.1037/0022-0663.85.3.478

Amato, J., & Watkins, M. W. (2011). The predictive validity of CBM writing indices of eighth-grade students. *Journal of Special Education, 44*, 195–204. doi:10.1177/0022466909333516

Arcia, E., & Conners, C. K. (1998). Gender differences in ADHD? *Journal of Developmental and Behavioral Pediatrics, 19*, 77–83. doi:10.1097/00004703-199804000-00003

Bereiter, C., & Scardamalia, M. (1987). *The psychology of written composition*. Hillsdale, NJ: Erlbaum.

Berman, R., & Verhoeven, L. (2002). Cross-linguistic perspectives on the development of text-production abilities. *Written Language and Literacy, 5*, 1–43. doi:10.1075/wll.5.1.02ber

Berninger, V. W. (1999). Coordinating transcription and text generation in working memory during composing: Automatized and constructive processes. *Learning Disability Quarterly, 22*, 99–112. doi:10.2307/1511269

Berninger, V. W., & Abbott, R. D. (2010). Listening comprehension, oral expression, reading comprehension, and written expression: Related yet unique language systems in Grades 1, 3, 5, and 7. *Journal of Educational Psychology, 102*, 635–651. doi:10.1037/a0019319

Berninger, V. W., Abbott, R. D., Abbott, S. P., Graham, S., & Richards, T. (2002). Writing and reading: Connections between language by hand and language by eye. *Journal of Learning Disabilities, 35*, 39–56. doi:10.1177/002221940203500104

Berninger, V. W., Abbott, R. D., Jones, J., Wolf, B. J., Gould, L., Anderson-Youngstrom, M., . . . Apel, K. (2006). Early development of language by hand: Composing, reading, listening, and speaking connections: Three-letter writing modes; and fast mapping in spelling. *Developmental Neuropsychology, 29*, 61–92. doi:10.1207/s15326942dn2901_5

Berninger, V. W., Abbott, R. D., Trivedi, P., Olson, E., Gould, L., Hiramatsu, S., . . . Westhagen, S. Y. (2010). Applying multiple dimensions of reading fluency to assessment and instruction. *Journal of Psychoeducational Assessment, 28*, 3–18. doi:10.1177/0734282909336083

Berninger, V. W., & Fuller, F. (1992). Gender differences in orthographic, verbal, and compositional fluency: Implications for assessing writing disabilities in primary grade children. *Journal of School Psychology, 30*, 363–382. doi:10.1016/0022-4405(92)90004-O

Berninger, V. W., & Swanson, H. L. (1994). Modifying Hayes and Flower's model of skilled writing to explain beginning and developing writing. In E. Butterfield (Ed.), *Children's writing: Toward a process theory of development of skilled writing* (pp. 57–81). Greenwich, CT: JAI Press.

Berninger, V. W., Vaughn, K. B., Graham, S., Abbott, R. D., Abbott, S. P., Rogan, L. W., . . . Reed, E. (1997). Treatment of handwriting problems in beginning writers: Transfer from handwriting to composition. *Journal of Educational Psychology, 89*, 652–666. doi:10.1037/0022-0663.89.4.652

Berninger, V. W., & Winn, W. D. (2006). Implications of advancements in brain research and technology for writing development, writing instruction, and educational evolution. In C. MacArthur, S. Graham, & J. Fitzgerald (Eds.), *Handbook of writing research* (pp. 96–114). New York, NY: Guilford Press.

Bollen, K. A. (1989). *Structural equations with latent variables*. New York, NY: Wiley.

Bowers, P. G. (1995). Tracing symbol naming speed's unique contributions to reading disabilities over time. *Reading and Writing, 7*, 189–216. doi:10.1007/BF01027185

Bruning, R., & Horn, C. (2000). Developing motivation to write. *Educational Psychologist, 35*, 25–37. doi:10.1207/S15326985EP3501_4

Carrow-Woolfolk, E. (2011). *Oral and Written Language Scales* (2nd ed.). Torrance, CA: Western Psychological Services.

Casas, A. M., Ferrer, M. S., & Fortea, I. B. (2013). Written composition performance of students with attention-deficit/hyperactivity disorder. *Applied Psycholinguistics, 34*, 443–460. doi:10.1017/S0142716411000828

Chan, D. W., Ho, C. S. H., Tsang, S. M., Lee, S. H., & Chung, K. K. H. (2006). Exploring the reading-writing connection in Chinese children with dyslexia in Hong Kong. *Reading and Writing, 19*, 543–561. doi:10.1007/s11145-006-9008-z

Chenoweth, N. A., & Hayes, J. R. (2003). The inner voice in writing. *Written Communications, 20*, 99–118. doi:10.1177/0741088303253572

Coker, D. L., & Ritchey, K. D. (2010). Curriculum-based measurement of writing in kindergarten and first grade: An investigation of production and qualitative scores. *Exceptional Children, 76*, 175–193. doi:10.1177/001440291007600203

Compton, D. L., DeFries, J. C., & Olson, R. K. (2001). Are RAN and phonological awareness deficits additive in children with reading disabilities? *Dyslexia, 7*, 125–149. doi:10.1002/dys.198

Cragg, L., & Nation, K. (2006). Exploring written narrative in children with poor reading comprehension. *Educational Psychology, 26*, 55–72. doi:10.1080/01443410500340991

de Jong, P. F., & van der Leij, A. (2003). Developmental changes in the manifestation of a phonological deficit in dyslexic children learning to read a regular orthography. *Journal of Educational Psychology, 95*, 22–40. doi:10.1037/0022-0663.95.1.22

Deno, S. L. (1985). Curriculum-based measurement: The emerging alternative. *Exceptional Children, 52*, 219–232.

Ding, Y., Richman, L. C., Yang, L., & Guo, J. (2010). Rapid automatized naming and immediate memory functions in Chinese Mandarin-speaking elementary readers. *Journal of Learning Disabilities, 43*, 48–61. doi:10.1177/0022219409345016

Dockrell, J. E., & Connelly, V. (in press). The role of oral language in underpinning the text generation difficulties in children with specific language impairment. *Journal of Research in Reading*.

Dockrell, J. E., Lindsay, G., & Connelly, V. (2009). The impact of specific language impairment on adolescents' written text. *Exceptional Children, 75*, 427–446. doi:10.1177/001440290907500403

Espin, C., Shin, J., Deno, S. L., Skare, S., Robinson, S., & Benner, B. (2000). Identifying indicators of written expression proficiency for middle school students. *Journal of Special Education, 34*, 140–153. doi:10.1177/002246690003400303

Espin, C. A., Weissenburger, J. W., & Benson, B. J. (2004). Assessing the writing performance of students in special education. *Exceptionality, 12*, 55–66. doi:10.1207/s15327035ex1201_5

Gansle, K. A., VanDerHeyden, A. M., Noell, G. H., Resetar, J. L., & Williams, K. L. (2006). The technical adequacy of curriculum-based and rating-based measures of written expression for elementary school students. *School Psychology Review, 35*, 435–450.

Gillam, R. B., & Pearson, N. A. (2004). *Test of Narrative Language*. Austin, TX: Pro-Ed.

Graham, S. (1990). The role of production factors in learning disabled students' compositions. *Journal of Educational Psychology, 82*, 781–791. doi:10.1037/0022-0663.82.4.781

Graham, S., Berninger, V. W., Abbott, R. D., Abbott, S. P., & Whitaker, D. (1997). Role of mechanics in composing of elementary school students: A new methodological approach. *Journal of Educational Psychology, 89*, 170–182. doi:10.1037/0022-0663.89.1.170

Graham, S., Berninger, V. W., & Fan, W. (2007). The structural relationship between writing attitude and writing achievement in first and third grade students. *Contemporary Educational Psychology, 32*, 516–536. doi:10.1016/j.cedpsych.2007.01.002

Graham, S., Harris, K. R., & Chorzempa, B. F. (2002). Contribution of spelling instruction to the spelling, writing, and reading of poor spellers.

Journal of Educational Psychology, 94, 669–686. doi:10.1037/0022-0663.94.4.669

Graham, S., Harris, K., & Fink, B. (2000). Is handwriting casually related to learning to write? Treatment of handwriting problems in beginning writers. *Journal of Educational Psychology*, 92, 620–633. doi:10.1037/0022-0663.92.4.620

Graham, S., Harris, K. R., & Hebert, M. (2011). *Informing writing: The benefits of formative assessment*. Washington, DC: Alliance for Excellent Education.

Graham, S., Harris, K. R., & Mason, L. (2005). Improving the writing performance, knowledge, and self-efficacy of struggling young writers: The effects of self-regulated strategy development. *Contemporary Educational Psychology*, 30, 207–241. doi:10.1016/j.cedpsych.2004.08.001

Gregg, N., Coleman, C., Stennett, R. B., & Davis, M. (2002). Discourse complexity of college writers with and without disabilities: A multidimensional analysis. *Journal of Learning Disabilities*, 35, 23–38. doi:10.1177/002221940203500103

Hammill, D. D., & Larsen, S. C. (1996). *Test of Written Language* (3rd ed.). Austin, TX: Pro-Ed.

Hammill, D. D., & Page, S. C. (2009). *Test of Written Language* (4th ed.). Austin, TX: Pro-Ed.

Hawke, J. L., Olson, R. K., Willcutt, E. G., Wadsworth, S. J., & DeFries, J. C. (2009). Gender ratios for reading difficulties. *Dyslexia*, 15, 239–242. doi:10.1002/dys.389

Hayes, J. (1996). A new framework for understanding cognition, and affect in writing. In M. Levy & S. Ransdell (Eds.), *The science of writing: Theories, methods, individual differences, and applications* (pp. 1–27). Mahwah, NJ: Erlbaum.

Hayes, J. R. (2012). Evidence from language bursts, revisions, and transcription for translation and its relation to other writing processes. In M. Fayol, D. Alamargot, & V. Berninger (Eds.), *Translation of thought to written text while composing: Advancing theory, knowledge, methods, and applications* (pp. 45–67). East Sussex, UK: Psychology Press.

Hidi, S., & Boscolo, P. (2006). Motivation and writing. In C. A. MacArthur, S. Graham, & J. Fitzgerald (Eds.), *Handbook of writing research* (pp. 144–157). New York, NY: Guilford Press.

Hooper, S. R., Costa, L. -J., McBee, M., Anderson, K. L., Yerby, D. C., Knuth, S. B., & Childress, A. (2011). Concurrent and longitudinal neuropsychological contributors to written language expression in first and second grade students. *Reading and Writing*, 24, 221–252. doi:10.1007/s11145-010-9263-x

Hooper, S. R., Swartz, C. W., Wakely, M. B., de Kruif, R. E. L., & Montgomery, J. W. (2002). Executive functions in elementary school children with and without problems in written expression. *Journal of Learning Disabilities*, 35, 57–68. doi:10.1177/002221940203500105

Jones, D., & Christensen, C. (1999). Relationship between automaticity in handwriting and students' ability to generate written text. *Journal of Educational Psychology*, 91, 44–49. doi:10.1037/0022-0663.91.1.44

Juel, C., Griffith, P. L., & Gough, P. B. (1986). Acquisition of literacy: A longitudinal study of children in first and second grade. *Journal of Educational Psychology*, 78, 243–255. doi:10.1037/0022-0663.78.4.243

Kail, R., & Hall, L. K. (1994). Processing speed, naming speed, and reading. *Developmental Psychology*, 30, 949–954. doi:10.1037/0012-1649.30.6.949

Kent, S., Wanzeck, J., Petscher, Y., Al Otaiba, S., & Kim, Y.-S. (in press). Writing fluency and quality in kindergarten and first grade: The role of attention, reading, transcription, and oral language. *Reading and Writing*.

Kim, Y.-S. (2011). Considering linguistic and orthographic features in early literacy acquisition: Evidence from Korean. *Contemporary Educational Psychology*, 36, 177–189.

Kim, Y.-S., Al Otaiba, S., Folsom, J. S., & Greulich, L. (2013). Language, literacy, attentional behaviors, and instructional quality predictors of written composition for first graders. *Early Childhood Research Quarterly*, 28, 461–469. doi:10.1016/j.ecresq.2013.01.001

Kim, Y.-S., Al Otaiba, S., Sidler, J. F., Greulich, L., & Puranik, C. (2014). Evaluating the dimensionality of first-grade written composition. *Journal of Speech, Language, and Hearing Research*, 57, 199–211.

Kim, Y.-S., Al Otaiba, S., Puranik, C., Folsom, J. S., Greulich, L., & Wagner, R. K. (2011). Componential skills of beginning writing: An exploratory study at the end of kindergarten. *Learning and Individual Differences*, 21, 517–525. doi:10.1016/j.lindif.2011.06.004

Kim, Y.-S., Park, C., & Park, Y. (2013). Is academic language use a separate dimension in beginning writing? Evidence from Korean children. *Learning and Individual Differences*, 27, 8–15. doi:10.1016/j.lindif.2013.06.002

Kim, Y.-S., Puranik, C., & Al Otaiba, S. (in press). Developmental trajectories of writing skills in first grade: Examining the effects of SES and language and/or speech impairments. *Elementary School Journal*.

Kirby, J., Parrila, R. K., & Pfeiffer, S. L. (2003). Naming speed and phonological awareness as predictors of reading development. *Journal of Educational Psychology*, 95, 453–464. doi:10.1037/0022-0663.95.3.453

Kline, R. B. (2005). *Principles and practice of structural equation modeling* (2nd ed.). New York, NY: Guilford Press.

Knudson, R. E. (1992). Development and application of a writing attitude survey for Grades 1 to 3. *Psychological Reports*, 70, 711–720. doi:10.2466/pr0.1992.70.3.711

Knudson, R. E. (1995). Writing experiences, attitudes, and achievement of first to sixth graders. *Journal of Educational Research*, 89, 90–97. doi:10.1080/00220671.1995.9941199

Kulikowich, J. M., Mason, L. H., & Brown, S. B. (2008). Evaluating fifth- and sixth-grade students' expository writing: Task development, scoring, and psychometric issues. *Reading and Writing*, 21, 153–175. doi:10.1007/s11145-007-9068-8

LaBerge, D., & Samuels, J. (1974). Towards a theory of automatic information processing in reading. *Cognitive Psychology*, 6, 293–323. doi:10.1016/0010-0285(74)90015-2

Lee, J. (2013). Can writing attitudes and learning behavior overcome gender difference in writing? Evidence from NAEP. *Written Communication*, 30, 164–193. doi:10.1177/0741088313480313

Lembke, E., Deno, S. L., & Hall, K. (2003). Identifying an indicator of growth in early writing proficiency for elementary school students. *Assessment for Effective Intervention*, 28, 23–35. doi:10.1177/073724770302800304

Levy, F., Hay, D. A., Bennett, K. S., & McStephen, M. (2005). Gender differences in ADHD subtype comorbidity. *Journal of the American Academy of Child & Adolescent Psychiatry*, 44, 368–376. doi:10.1097/01.chi.0000153232.64968.c1

Limpo, T., & Alves, R. A. (2013). Modeling writing development: Contribution of transcription and self-regulation to Portuguese students' text generation quality. *Journal of Educational Psychology*, 105, 401–413. doi:10.1037/a0031391

Mackie, C., & Dockrell, J. E. (2004). The nature of written language deficits in children with SLI. *Journal of Speech, Language, and Hearing Research*, 47, 1469–1483. doi:10.1044/1092-4388(2004/109)

McCutchen, D., Francis, M., & Kerr, S. (1997). Revising for meaning: Effects of knowledge and strategy. *Journal of Educational Psychology*, 89, 667–676.

McGrew, K. S., Schrank, F. A., & Woodcock, R. W. (2007). *Technical manual: Woodcock-Johnson III Normative Update*. Rolling Meadows, IL: Riverside.

McKenna, M., Kear, D., & Ellsworth, R. (1995). Children's attitudes toward reading: a national survey. *Reading Research Quarterly*, 30, 934–956.

McMaster, K. L., Du, X., & Pétursdóttir, A. L. (2009). Technical features of curriculum-based measures for beginning writers. *Journal of Learning Disabilities*, 42, 41–60. doi:10.1177/0022219408326212

McMaster, K. L., Du, X., Yeo, S., Deno, S. L., Parker, D., & Ellis, T. (2011). Curriculum-based measures of beginning writing: Technical features of the slope. *Exceptional Children*, 77, 185–206. doi:10.1177/001440291107700203

McMaster, K., & Espin, C. (2007). Technical features of curriculum-based measurement in writing: A literature review. *Journal of Special Education*, 41, 68–84. doi:10.1177/00224669070410020301

Miles, T. R., Haslum, M. N., & Wheeler, T. J. (1998). Gender ratio in dyslexia. *Annals of Dyslexia*, 48, 27–55. doi:10.1007/s11881-998-0003-8

Muthén, L. K., & Muthén, B. O. (2012). *Mplus* (Version 7) [Computer software]. Los Angeles, CA: Muthén and Muthén.

National Center for Education Statistics. (2011). *The Nation's Report Card: Reading 2011* (NCES 2012–457). Washington, DC: Author.

Northwest Regional Educational Laboratory. (2011). *6 + 1 trait writing*. Retrieved from <http://educationnorthwest.org/traits>

Olinghouse, N. G. (2008). Student- and instruction-level predictors of narrative writing in third-grade students. *Reading and Writing*, 21, 3–26. doi:10.1007/s11145-007-9062-1

Olinghouse, N. G., & Graham, S. (2009). The relationship between discourse knowledge and the writing performance of elementary-grade students. *Journal of Educational Psychology*, 101, 37–50. doi:10.1037/a0013462

Pajares, F. (2003). Self-efficacy beliefs, motivation, and achievement in writing: A review of the literature. *Reading & Writing Quarterly*, 19, 139–158. doi:10.1080/10573560308222

Pajares, F., & Valiante, G. (1999). Grade level and gender differences in the writing self-beliefs of middle school students. *Contemporary Educational Psychology*, 24, 390–405. doi:10.1006/ceps.1998.0995

Persky, H. R., Dane, M. C., & Jin, Y. (2003). *The Nation's Report Card: Writing 2002* (NCES 2003–529). Washington, DC: National Center for Education Statistics. Retrieved from http://nces.ed.gov/nations_reportcard

Posner, M. I., & Snyder, C. R. R. (1975). Attention and cognition control. In R. Solo (Ed.), *Information processing and cognition: The Loyola Symposium* (pp. 55–85). Hillsdale, NJ: Erlbaum.

Puranik, C. S., Lombardino, L. J., & Altmann, L. J. (2007). Writing through retellings: An exploratory study of language-impaired and dyslexic populations. *Reading and Writing*, 20, 251–272. doi:10.1007/s11145-006-9030-1

Puranik, C. S., Lombardino, L., & Altmann, L. (2008). Assessing the microstructure of written language using a retelling paradigm. *American Journal of Speech Language Pathology*, 17, 107–120. doi:10.1044/1058-0360(2008/012)

Re, A. M., Pedron, M., & Cornoldi, C. (2007). Expressive writing. Difficulties in children described as exhibiting ADHD symptoms. *Journal of Learning Disabilities*, 40, 244–255. doi:10.1177/00222194070400030501

Ritchey, K. D., McMaster, K. L., Al Otaiba, S., Puranik, C. S., Kim, Y.-S., Parker, D. C., & Ortiz, M. (in press). Indicators of fluent writing in beginning writers. In K. Cummings & Y. Petscher (Eds.), *The fluency construct*. New York, NY: Springer.

Sáez, L., Folsom, J. S., Al Otaiba, S., & Schatschneider, C. (2012). Relations among student attention behaviors, teacher practices, and beginning word reading skill. *Journal of Learning Disabilities*, 45, 418–432. doi:10.1177/0022219411431243

Savage, R. S., Frederickson, N., Goodwin, R., Patni, U., Smith, N., & Tuersley, L. (2005). Relationships among rapid digit naming, phonological processing, motor automaticity, and speech perception in poor, average, and good readers and spellers. *Journal of Learning Disabilities*, 38, 12–28. doi:10.1177/00222194050380010201

Scardamalia, M., Bereiter, C., & Goleman, H. (1982). The role of production factors in writing ability. In M. Nystrand (Ed.), *What writers know: The language, process, and structure of written discourse* (pp. 175–210). San Diego, CA: Academic Press.

Scott, C., & Windsor, J. (2000). General language performance measures in spoken and written discourse produced by school-age children with and without language learning disabilities. *Journal of Speech, Language, and Hearing Research*, 43, 324–339.

Shanahan, T. (2006). Relations among oral language, reading, and writing development. In C. A. MacArthur & S. Graham (Eds.), *Handbook of writing* (pp. 171–183). New York, NY: Guilford Press.

Shanahan, T., & Lomax, R. G. (1986). An analysis and comparison of theoretical models of the reading-writing relationship. *Journal of Educational Psychology*, 78, 116–123. doi:10.1037/0022-0663.78.2.116

Shaywitz, S. E., Shaywitz, B. A., Fletcher, J. M., & Escobar, M. D. (1990). Prevalence of reading disability in boys and girls. *JAMA: Journal of the American Medical Association*, 264, 998–1002. doi:10.1001/jama.1990.03450080084036

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlation: Uses in assessing rater reliability. *Psychological Bulletin*, 86, 420–428. doi:10.1037/0033-2909.86.2.420

Spring, C., & Davis, J. M. (1988). Relations of digit naming speed with three components of reading. *Applied Psycholinguistics*, 9, 315–334. doi:10.1017/S0142716400008031

Swanson, J. M., Schuck, S., Mann, M., Carlson, C., Hartman, K., Sergeant, J. A., . . . McCleary, R. (2006). *Categorical and dimensional definitions and evaluations of symptoms of ADHD: The SNAP and SWAN Rating Scales*. Retrieved from http://www.adhd.net/SNAP_SWAN.pdf

Torgesen, J. K., Wagner, R. K., & Rashotte, C. A. (2012). *Test of Word Reading Efficiency* (2nd ed.). Austin, TX: Pro-ED.

Wagner, R. K., Puranik, C. S., Foorman, B., Foster, E., Tschinkel, E., & Kantor, P. T. (2011). Modeling the development of written language. *Reading and Writing*, 24, 203–220. doi:10.1007/s11145-010-9266-7

Wagner, R. K., & Torgesen, J. K. (1987). The nature of phonological processing and its causal role in the acquisition of reading skills. *Psychological Bulletin*, 101, 192–212. doi:10.1037/0033-2909.101.2.192

Wagner, R. K., Torgesen, J. K., Rashotte, C. A., & Pearson, N. A. (2010). *Test of Silent Reading Efficiency and Comprehension*. Austin, TX: Pro-ED.

Wechsler, D. (2009). *Wechsler Individual Achievement Test* (3rd ed.). San Antonio, TX: Pearson.

Wolf, M., & Bowers, P. (1999). The double-deficit hypothesis for the developmental dyslexias. *Journal of Educational Psychology*, 91, 415–438. doi:10.1037/0022-0663.91.3.415

Wolf, M., & Denckla, M. B. (2005). *RAN/RAS: Rapid Automatized Naming and Rapid Alternating Stimulus Tests*. Austin, TX: Pro-ED.

Wolf, M., & O'Brien, B. (2001). On issues of time, fluency, and intervention. In A. Fawcett (Ed.), *Dyslexia: Theory and good practice* (pp. 124–140). London, UK: Whurr.

Woodcock, R. W., McGrew, K. S., & Mather, N. (2001). *Woodcock-Johnson III Tests of Achievement*. Itasca, IL: Riverside.

Yan, C. M. W., McBride-Chang, C., Wagner, R. K., Zhang, J., Wong, A. M. Y., & Shu, H. (2012). Writing quality in Chinese children: Speed and fluency matter. *Reading and Writing*, 25, 1499–1521.

Yoshimasu, K., Barbaresi, W. J., Colligan, R. C., Killian, J. M., Voigt, R. G., Weaver, A. L., & Katusic, S. K. (2010). Gender, attention-deficit/hyperactivity disorder, and reading disability in a population-based birth cohort. *Pediatrics*, 126, e788–e795.

Cross-Language Transfer of Word Reading Accuracy and Word Reading Fluency in Spanish–English and Chinese–English Bilinguals: Script-Universal and Script-Specific Processes

Adrian Pasquarella
University of Delaware

Xi Chen
University of Toronto

Alexandra Gottardo
Wilfrid Laurier University

Esther Geva
University of Toronto

This study examined cross-language transfer of word reading accuracy and word reading fluency in Spanish–English and Chinese–English bilinguals. Participants included 51 Spanish–English and 64 Chinese–English bilinguals. Both groups of children completed parallel measures of phonological awareness, rapid automatized naming, word reading accuracy, and word reading fluency in their first language (L1) and in English, their second language (L2) in Grade 1. Word reading accuracy and word reading fluency were assessed in L1 and L2 again in Grade 2. Cross-language transfer of word reading accuracy was found only in the Spanish–English bilinguals. In contrast, cross-language transfer of word reading fluency was found in both the Spanish–English bilinguals and the Chinese–English bilinguals. Our results suggest transfer of word reading accuracy is based on the structural similarities between the L1 and L2 scripts. By contrast, word reading fluency operates largely as a script-universal process. Implications for reading theory and for assessment and instruction of bilingual children are discussed.

Keywords: cross-language transfer, word reading accuracy, word reading fluency, bilingual, English language learners

Supplemental materials: <http://dx.doi.org/10.1037/a0036966.supp>

Both word reading accuracy and word reading fluency are critical for successful comprehension (Ehri, 1997; Gough & Tunmer, 1986; Hoover & Gough, 1990; LaBerge & Samuels, 1974; Perfetti & Hogaboam, 1975; Stanovich, 1991). Accurate word recognition leads to correct lexical activation (Perfetti, 1985; Stanovich, 1991), whereas fluent word recognition is critical for rapid processing of orthographic information. Together, accurate and efficient lexical access allows for greater capacity for higher level processing such as constructing meaning at the sentence, paragraph, and discourse levels of a text (Breznitz, 2006; Perfetti, 1985; Swanson & Berninger, 1995; Stanovich, 1994).

Research has demonstrated that for bilingual children, many skills developed in their first language (L1) are positively related to reading acquisition in their second language (L2; e.g., Durgunoğlu, 2002; Genesee, Geva, Dressler, & Kamil, 2006;

Gottardo, 2002). The present study focused on cross-language transfer of word reading accuracy and word reading fluency in bilingual children. The transfer patterns of these two constructs, especially those of word reading fluency, have not been systematically examined. Notably, there is little agreement in the literature as to what constitutes transfer (Koda, 2007). Traditionally, transfer is defined as the use of linguistic (and cognitive) knowledge acquired in L1 for L2 learning (Odlin, 1989). This type of transfer hinges upon the linguistic distance between L1 and L2 (Koda, 2007). When the L1 and L2 are closely related, shared structural properties pose similar demands on processing and allow L1 competencies to function in L2 reading with little adjustment. By contrast, L1 skills do not facilitate L2 reading to the same extent when the two languages are distantly related. Recently, transfer has also been conceptualized as the ability to develop L2 proficiency by drawing on previously acquired resources (Genesee et al., 2006; Koda, 2007). Although these resources, such as phonological awareness and verbal working memory, are first acquired in bilingual children's L1, they are important for developing reading skills in any language and are thereby considered script universal (Abu-Rabia & Siegel, 2002; Da Fontoura & Siegel, 1995; Genesee & Geva, 2006).

Our study presents a unique opportunity to distinguish between the two types of cross-language transfer. Unlike most previous transfer studies that examined bilinguals only from a single language background, this study simultaneously involved two groups

This article was published Online First June 2, 2014.

Adrian Pasquarella, School of Education, University of Delaware; Xi Chen, Ontario Institute for Studies in Education, University of Toronto; Alexandra Gottardo, Department of Psychology, Wilfrid Laurier University; Esther Geva, Ontario Institute for Studies in Education, University of Toronto.

Correspondence concerning this article should be addressed to Adrian Pasquarella, University of Delaware, School of Education, Willard Hall, Newark, DE 19716. E-mail: a.pasquarella@gmail.com

of bilinguals, Spanish–English bilinguals and Chinese–English bilinguals. We compared cross-language transfer of two aspects of reading at the word level: word reading accuracy and word reading fluency. The L1s of the two groups, Spanish and Chinese, have many contrasting features in terms of how the oral language is represented by the script, as well as in terms of how much overlap it has with English, the children's L2. Specifically, Spanish and English share the same alphabetic script and the use of grapheme-to-phoneme correspondences in reading, while Chinese is a logographic script that emphasizes the mapping between characters and morphemes. Thus, comparing transfer patterns in these two bilingual groups can elucidate whether and how L1 characteristics influence patterns of transfer and reveal script-universal and script-specific processes in cross-language transfer.

Word Reading Accuracy

Word reading accuracy refers to the ability to accurately identify single words from print. Critically important for reading comprehension, it is a skill that all children must master in the early grades. Word reading accuracy is typically assessed by asking children to read printed words out loud. A fundamental and universal aspect of reading is that every orthography encodes the spoken language using abstract symbols. Reading a word accurately in any language requires linkages between orthographic, phonological, and meaning representations (Seidenberg & McClelland, 1989). However, the way that orthography maps onto phonology and morphology differs across languages and influences the reading process (Perfetti, 2003; Ziegler & Goswami, 2005).

Both English and Spanish are alphabetic languages based on the Roman script. However, the two scripts differ with respect to the transparency of grapheme-to-phoneme correspondences. English is considered a deep orthography because it has many-to-many sound-to-symbol correspondences (e.g., *ch* in *chord*, *chore*, *chute*; *ea* in *heal* vs. *healthy*) (Geva & Siegel, 2000; Gholamain & Geva, 1999; Venezky, 1970). Spanish, on the other hand, is a shallow orthography with near perfect grapheme-to-phoneme correspondences (Jímenez-González, 1997). Despite the differences in grapheme–phoneme correspondences, beginning readers of both English and Spanish rely heavily on phonologically based skills, such as phonological awareness and decoding, to read words (Durgunoğlu, Nagy, & Hancin-Bhatt, 1993; Lindsey, Manis, & Bailey, 2003). In fact, research has shown that phonological skills are related to reading ability and disability for children who learn to read Spanish or English, as well as other alphabetic languages (Ball & Blachman 1991; Shaywitz & Shaywitz, 2005; Stanovich, 1988; Torgesen, Wagner, Rashotte, Rose, et al., 1999).

The Chinese orthography is logographic in nature, with each Chinese character corresponding to both a morpheme and a syllable in the spoken language. For example, the character 兵 means *soldier* and is pronounced /bing1/. The majority of Chinese characters are phonetic compound characters. Each compound contains a phonetic component, which may or may not provide useful information about the pronunciation of the character. For example, 清 /qing1/ is a regular character in which the phonetic 青 /qing1/ has the same pronunciation as the compound character; 倩 /qian4/ is an irregular character in which the pronunciation of the phonetic 青 is misleading. By the end of Grade 6, Chinese children are required to learn about 2,500 characters, and 72% of them are

phonetic compounds (Shu, Chen, Anderson, Wu, & Xuan, 2003). The phonetic component only accurately represents character pronunciation in 39% of the compounds (Shu et al., 2003).

There are two ways to read Chinese characters. A character without an internal structure, such as 兵, can only be read by mapping the whole character to its pronunciation. A compound character can be read either by mapping the whole character to its pronunciation or by naming the phonetic, although the latter is not always reliable. Thus, reading Chinese requires mapping orthographic patterns to sound, which is different from the decoding strategy used for reading alphabetic languages. Due to the characteristics of the Chinese orthography, although phonological skills are related to reading in young Chinese children (e.g., McBride-Chang & Ho, 2005), other metalinguistic skills such as morphological awareness and orthographic processing account for more variance in Chinese reading than phonological skills (e.g., Keung & Ho, 2009; Leong, Tse, Lon, & Hau, 2008; Liao, Georgiou, & Parrila, 2008; McBride-Chang et al., 2005; Tan & Perfetti, 1998; Tong & McBride-Chang, 2010; Yeung et al., 2011). Therefore, based on script characteristics, somewhat different processes are involved in reading the L1s of Spanish–English and Chinese–English bilinguals.

A growing number of studies have provided evidence for cross-language transfer of word reading accuracy between two alphabetic scripts (Gholamain & Geva 1999; Gottardo, 2002; Manis, Lindsey, & Bailey, 2004; Páez & Rinaldi, 2006). For example, in a large-scale longitudinal study, Lindsey et al. (2003) found that for Spanish-speaking English language learners, Spanish phonological awareness and Spanish word reading accuracy measured in kindergarten were both predictive of English word identification in Grade 1. Based on their review of previous research, Dressler and Kamil (2006) concluded that word reading skills transfer across two alphabetic languages whether they are structurally close (e.g., Spanish–English) or distant (e.g., Arabic–English). They also pointed out that the transfer occurs in L2 readers with a wide range of ages and proficiency levels.

Studies of cross-language transfer of word reading skills between Chinese and English, however, have produced mixed results. Gottardo, Yan, Siegel, and Wade-Woolley (2001) found that in Chinese–English bilinguals (mean age of approximately 10 years old), word reading skills in Chinese and English were not significantly correlated with one another, although phonological skills were correlated across the two languages. Similar findings were reported in Wang, Perfetti, and Liu (2005) for Chinese–English bilinguals in Grades 2 and 3. The findings of these two studies, together with those involving Spanish–English bilinguals, seem to suggest that transfer of word reading accuracy is constrained by similarities between bilingual children's L1 and L2. However, in another study, Keung and Ho (2009) observed a significant correlation between Chinese and English word reading skills among Grade 2 children in Hong Kong.

For Chinese–English readers, the difference in the findings may be attributed to the educational context. The participants of both Gottardo et al. (2001) and Wang et al. (2005) were Chinese–English bilinguals in North America, where English is the societal language and the language of instruction. North American children are likely to use phonological and decoding strategies to read English because phonics instruction is a major component in early reading programs. Cantonese was the medium of instruction for

the Hong Kong children in Keung and Ho (2009). These children were exposed to English mainly in English language classes, where phonics instruction was not provided. Thus, it is plausible that the participants of Keung and Ho (2009) applied L1 strategies (i.e., whole word activation) to reading English, which strengthened the cross-language connection between English and Chinese word reading. Given the inconsistency, more research is needed to investigate the nature of cross-language transfer of reading skills in Chinese–English bilinguals.

Word Reading Fluency

Reading fluency is “the oral translation of text with speed and accuracy” (Fuchs, Fuchs, Hosp, & Jenkins, 2001, p. 239). In beginning readers who are not yet able to read connected text, reading fluency is often assessed by the rate and accuracy of reading lists of isolated words or pseudowords aloud (Good, Simmons, & Kame’enui, 2001; Jenkins, Fuchs, van den Broek, Espin, & Deno, 2003). A typical word reading fluency measure instructs children to read aloud as many words or pseudowords as possible within a specified time (Torgesen, Wagner, & Rashotte, 1999). Research has shown that word reading fluency is an effective screening measure for determining at-risk status among beginning readers (Clemens, Shapiro, & Thoemmes, 2011; Compton et al., 2010; Compton, Fuchs, Fuchs, & Bryant, 2006; Fuchs, Fuchs, & Compton, 2004). Furthermore, word reading fluency measured at an early age is a strong predictor of later text reading fluency (Geva & Farnia, 2012; Good et al., 2001; Jenkins et al., 2003) and reading comprehension (Aaron, Joshi, & Williams, 1999; Fuchs, Fuchs, & Maxwell, 1988; Joshi & Aaron, 2000; Marston, 1989; National Institute of Child Health and Human Development, 2000). Good et al. (2001) showed that for children in Grade 1, word reading fluency measured with pseudowords in the first semester was highly correlated with text reading fluency as well as reading performance in the second semester. Geva and Farnia (2012) reported that word and text reading fluency loaded on a single factor for both English L1 and L2 students in Grade 2. Thus, there is a high degree of overlap between word and text reading fluency in the early grades, and word reading fluency may form the foundation for developing text reading fluency. In the present study, we chose to measure reading fluency with isolated words because our participants were bilingual children in Grade 1 at the beginning of the study and would have had difficulty reading connected text.

Word reading fluency is characterized by rapid, automatic word recognition (Kuhn, Schwanenflugel, & Meisinger, 2010; Samuels, 2006). According to Kuhn et al. (2010), automaticity possesses four features: speed, effortlessness, autonomy, and lack of conscious awareness. A fluent reader reads by retrieving words directly from long-term memory, as opposed to by phonological recoding, which is slower and more laborious. In addition, rapid word recognition occurs without intention (i.e., autonomy) or conscious awareness of the component skills (e.g., phonological awareness, morphological awareness, orthographic processing) required for reading. Reading comprehension demands substantial cognitive resources. It is a challenging task for beginning readers because they must allocate a large amount of cognitive resources to word recognition. When word recognition becomes more fluent, readers can reallocate resources to text processing and achieve

better comprehension (Fuchs et al., 2004; Perfetti, 1985; Stanovich, 1980, 1994; Swanson & Berninger, 1995).

Although learning to read English and Chinese requires somewhat different component skills, in both orthographies, fluent reading is developed by creating strong links among orthographic, phonological, and semantic patterns following repeated exposures to print (Clemens et al., 2011; Kuhn et al., 2010; Seidenberg, 2005; Seidenberg & McClelland, 1989; Shu & Anderson, 1999). There is evidence that similar processes are involved in reading Spanish sight words even though it is a more transparent orthography (Defior, Cary, & Martos, 2002; Wimmer & Goswami, 1994). Additional evidence for the use of similar processes in fluent reading across languages is the existence of the word superiority effect in alphabetic languages as well as in Chinese (Mattingly & Xu, 1993; Morton, 1969; Reicher, 1969). Across languages, fluent readers are better at tasks involving words than strings of letters. Breznitz (2003, 2006; Breznitz & Berman, 2003) posited that rapid word reading results from successful and efficient synchronization and integration of phonological, orthographic, and semantic information. Because each of these components exists to some extent in written script, it is reasonable to assume that this process is similar across Spanish, Chinese, and English.

To our knowledge, only one previous study examined transfer of reading fluency. In a 1-year longitudinal study, De Ramírez and Shapiro (2007) investigated the relationship between text reading fluency in Spanish (L1) and English (L2) among bilingual students in Grades 1–5. They found that Spanish fluency in the fall was correlated with English fluency in the spring. This result provides preliminary evidence for cross-language transfer of text reading fluency. However, the cross-language associations may have been due to spurious third variables because within-language measures known to be related to reading fluency, such as rapid naming and phonological awareness, were not controlled for in the analysis. In the present study, we extended previous research by including multiple within-language controls (e.g., nonverbal reasoning, phonological awareness, rapid naming, and an autoregressor of word reading fluency). In addition, we focused on word reading fluency as little is known about how this construct is related across bilingual children’s L1 and L2.

The Present Study

The present study examined cross-language transfer of word reading accuracy and fluency in Spanish–English and Chinese–English bilinguals. With respect to word reading accuracy, we predicted that cross-language transfer would vary as a function of similarities between the bilingual children’s L1 and L2. A stronger crossover effect would be observed between English and Spanish than between English and Chinese because English and Spanish are both alphabetic scripts that require grapheme-to-phoneme correspondences in reading. Whereas Spanish-speaking students bring to English reading alphabetic decoding abilities developed in their L1, the same cannot be said of Chinese-speaking students, as they use different strategies in L1 reading. With respect to word reading fluency, we predicted that cross-language transfer would occur for both Spanish–English and Chinese–English bilinguals. This prediction stems from the script-universal perspective of cross-language transfer. Considering that reading in any language requires efficient synchronization of phonological, orthographic, and

semantic information, reading fluency may be a script-universal construct that is related across bilingual children's L1 and L2 regardless of the degree of overlap between them.

The current study adopted a cross-lagged design. Participants were assessed twice, in Grades 1 and 2, respectively. Cross-lagged designs are more rigorous than cross-sectional correlational designs as they can account for the effect of an autoregressor, which is the outcome variable measured at an earlier time point. Without accounting for the autoregressor, relationships among the other predictors can be artificially inflated (Kenny, 1975). Additionally, we controlled for nonverbal reasoning, within-language phonological awareness, and rapid automatized naming (RAN) in our cross-language analysis. For instance, when English word reading accuracy in Grade 2 was the dependent variable, nonverbal reasoning, English phonological awareness, English RAN, and English word reading accuracy in Grade 1 were entered in the model before Grade 2 L1 word reading accuracy, the target predictor. Controlling for the autoregressor and multiple within-language variables greatly reduced the possibility that any cross-language transfer observed could be due to within-language relationships or other spurious third variables.

Method

Participants

Data collection occurred at two time points spaced 1 year apart, spring of Grade 1 and spring of Grade 2. In Grade 1, participants included 61 Spanish–English bilingual children and 68 Chinese–English bilingual children. One year later, 51 Spanish–English (mean age = 80.90 months, 30 girls) children and 64 Chinese–English (mean age = 81.43 months, 36 girls) remained in the study. Attrition occurred because some children moved away. Data were analyzed only for the students who participated in the study at both time points. The Chinese–English bilingual sample consisted of both Cantonese and Mandarin speakers. Of the 64 Chinese–English bilinguals who were used for data analysis, there were 49 Cantonese speakers and 15 Mandarin speakers. Because the Cantonese and Mandarin speakers performed similarly on most measures, the two subgroups were collapsed in data analysis to increase power. The only group difference between the Mandarin and Cantonese speakers was on phonological awareness in Chinese in Grade 1, with the Mandarin speakers performing significantly better on this measure.

The children were recruited from 19 schools in predominantly middle-class neighborhoods in two Canadian cities. English was the language of instruction in all schools. Approximately 70% of the Spanish–English bilinguals and 90% of the Chinese–English attended heritage language classes for 2.5 hr per week. Within Spanish–English homes, 30% of parents reported speaking only in Spanish. The other 70% reported conversations occurring in both Spanish and English. Within the Chinese–English homes, 35% of the sample reported speaking only in Chinese, with the other 65% reporting conversations occurring in both Chinese and English. The mean level of parental education was high school for both groups. Approximately 60% of the parents had a high school education or less, and 25% had a university education, whereas the other 15% did not report their level of education.

Measures

A battery of tests including nonverbal reasoning, phonological awareness, RAN of digits, word reading accuracy, and word reading fluency was given to each participant in Grade 1. All measures except for nonverbal reasoning were given in both English and the participant's L1. One year later, the same word reading accuracy and fluency measures were given in English and the participant's L1 in Grade 2.

Nonverbal reasoning. The Matrix Analogies Reasoning Test (Naglieri, 1985) was administered as a measure of nonverbal reasoning. This test involved presenting 64 pictures organized into four subtests of 16 items each. All participants attempted each of the four subtests. For each item, the child was shown a pattern with one portion missing and was asked to choose among six options the one that correctly completed the pattern. A stopping rule of four consecutive incorrect items was used for each subtest. The Cronbach's alpha for children aged 6 to 7 years old is .94 based on the testing manual (Naglieri, 1985).

Phonological awareness. English phonological awareness was assessed by an experimental deletion test (Jared, Cormier, Levy, & Wade-Woolley, 2011). The test consisted of three subtests assessing syllable, onset-rime, and phoneme deletion, respectively. The subtests were administered in that order, with five practice items and 12 test items in each subtest. The maximum score of the test was 36. In the syllable deletion subtest, for example, the child was asked to delete the first or second syllable from a multisyllable word: "Say *bam-daw*, now say what is left of *bam-daw* if you don't say *daw*." If the child scored two or fewer items correctly in a subtest, subsequent levels were not administered. The Cronbach's alpha calculated for our study was .86.

Spanish phonological awareness was measured by the Elision subtest of the Test of Phonological Processing in Spanish (TOPPS; Francis et al., 2001). The test contained three practice items and 20 test items (three syllable deletion items and 17 phoneme deletion items). The test was stopped when three consecutive incorrect responses were made. A parallel measure was used for Chinese–English bilinguals in Cantonese (Gottardo et al., 2001) and adapted for Mandarin. The Chinese measure contained the same numbers of practice and test items and followed the same procedure as the Spanish measure. The Cronbach's alpha calculated for our study was .88 for the Spanish measure and .95 for the Chinese measure.

Rapid Automatized Naming (RAN)–Digits. Rapid naming was assessed by the RAN–Digits subtest from the Comprehensive Test of Phonological Processing (Wagner, Torgesen, & Rashotte, 1999) in English and the RAN–Digits subtest of the TOPPS (Francis et al., 2001). A parallel experimental measure was adapted to Chinese. The same testing stimuli were used, but children were requested to respond either in Mandarin or in Cantonese. In all three measures, the child was presented with six rows of the same six digits arranged in different orders and was asked to name the 36 digits as quickly and accurately as possible. Two practice examples were given prior to testing to ensure that the child understood the instructions and was able to name the digits. Two alternative forms (A and B) were completed, and the time it took to name all the digits in seconds was recorded and used as the raw score in the analyses. The test–retest reliability for children aged 5 to 7 was .91 for the English measure according to the test manual (Wagner et al., 1999).

Word reading accuracy. Word reading accuracy was assessed by the Word Identification subtest of the Woodcock Reading Mastery Test—Revised (Woodcock, 1991) in English and by the *Identificación de letras y palabras* subtest of Woodcock Language Proficiency Battery—Revised, Spanish Form (Woodcock & Munoz-Sandoval, 1995) in Spanish. Both tests required the child to read words that increased in length and difficulty and had a stopping rule of six consecutive words read incorrectly. According to the manuals, the Cronbach's alpha was .92 for the English measure and .95 for the Spanish measure (Woodcock, 1991).

Word reading accuracy in Chinese was measured by a test previously used in Gottardo et al. (2001). The same set of 100 characters was used for both Cantonese and Mandarin speakers. The Cantonese items consisted of traditional characters, whereas the Mandarin items were in the simplified form. The items were presented in order of increasing difficulty. Children were encouraged to attempt all characters and were allowed to skip items or guess when they were unsure. The Cronbach's alpha calculated for this sample was .95.

Word reading fluency. Word reading fluency was assessed by the real-word subtest of the Test of Word Reading Efficiency (TOWRE) in English and the TOWRE words subtest from the TOPPS in Spanish (Francis et al., 2001). For both tests, raw scores were the number of items read correctly from a list of words in 45 s. According to the manuals, the test-retest reliability for ages 6 and 7 years was .97 (Torgesen, Wagner, & Rashotte, 1999) for the English version and .94 for the Spanish version. A parallel experimental measure following the same procedure was administered in Chinese. The test consisted of 104 items that were presented in order of increasing difficulty. Traditional characters were used for the Cantonese version, whereas simplified characters were used for the Mandarin version. The Cronbach's alpha calculated for this sample was .90.

Comparability of the Word Reading Measures

Comparability of measures used across languages was a central and critical aspect of the present study. It was necessary to establish that all parallel tasks were valid and reliable and that they measured equivalent constructs across languages. The English and Spanish tasks were taken from standardized assessment packages with established validity and high rates of reliability (Francis et al., 2001; Torgesen, Wagner, & Rashotte, 1999; Woodcock, 1991; Woodcock & Munoz-Sandoval, 1995). Since no standardized measures existed for Chinese, experimental measures were used in the present study. The Chinese measures were specifically designed to be analogous in the construction, instructional protocol, and difficulty level of their English and Spanish counterparts. Research based on the Chinese measures has been reported in many studies, and similar patterns of results occurred across these studies, suggesting that the measures were valid and had a high degree of reliability (see Gottardo et al., 2001; Leong, Cheng, & Mulcahy, 1987; Marinova-Todd, Zhao, & Bernhardt, 2010; Wang et al., 2005). Thus, the formats of English, Spanish, and Chinese measures were parallel across languages, and appropriately assessed the intended constructs.

As an additional check to ensure that the word reading accuracy and word reading fluency measures were comparable across the three languages, we calculated the mean frequency for each mea-

sure in each language. We used the SUBTLEX databases in English (Brysbaert & New, 2009), Spanish (Cuetos, Glez-Nosti, Barbón, & Brysbaert, 2011), and Chinese (Cai & Brysbaert, 2010) in our calculation, with the log frequency of a word's occurrence per million as the unit of analysis. The mean frequencies for the word reading accuracy measures were 3.42, 3.50, and 3.60 in English, Spanish, and Chinese, respectively. The mean frequencies for the word reading fluency measures were 3.05, 3.25, and 3.23 in English, Spanish, and Chinese, respectively. Univariate analysis of variance revealed no significant differences for either the word reading accuracy measures ($p = .429$) or the word reading fluency measures ($p = .575$), suggesting the levels of difficulty were similar across languages.

Procedure

All children were tested individually by trained research assistants in a quiet room during the school day. The children were tested in English and in their L1 on different days, with the order of L1 and L2 testing counterbalanced. The Spanish and Chinese measures were administered by research assistants who were native speakers of the respective languages. Both English and L1 instructions were used for the L1 measures to ensure that children understood the tasks. Only English instructions were used for the English measures. On average, it took the students approximately 1 hr to complete the measures in English and their L1.

Results

All measures were checked for normality, skewness, and kurtosis. Some measures were positively skewed (e.g., English and L1 RAN in Grade 1 for both groups, English and L1 word reading fluency in Grade 1 for both groups). For each of these measures, square root or logarithm transformations were performed to correct the distribution to normal. However, subsequent analyses conducted with transformed scores produced results virtually identical to those with untransformed scores. Therefore, the results with the untransformed raw scores are presented for all analyses.

Table 1 presents the mean raw scores, standard deviations, maximum scores, skewness, and kurtosis statistics for all measures for the Spanish–English and Chinese–English bilinguals. We conducted t tests on the English measures across groups. The two groups of bilinguals performed similarly on English phonological awareness and English RAN. The Chinese–English bilinguals outperformed the Spanish–English bilinguals on the English word reading accuracy and fluency measures in Grade 1; however, these differences became negligible by Grade 2. The results suggest that the Chinese–English and Spanish–English bilinguals had similar levels of English proficiency in Grade 2. A Pearson correlation matrix for all the variables is displayed by language group in Table 2. To protect against Type I error, only results significant at $p < .01$ were interpreted as meaningful. Overall, most variables were significantly correlated. Notably, scores on word reading fluency were significantly correlated across languages in both grades for both groups of participants with moderate to high correlations. For the Spanish–English bilinguals, scores on word reading accuracy were significantly correlated across languages in both grades. For the Chinese–English bilinguals, only Chinese word reading accuracy in Grade 1 was significantly correlated with English word reading accuracy in Grade 1.

Table 1

Descriptive Statistics of Chinese–English and Spanish–English Bilinguals

Variable	Chinese–English bilinguals					Spanish–English bilinguals				
	Maximum	<i>M</i>	<i>SD</i>	Skewness (<i>SE</i> = .30)	Kurtosis (<i>SE</i> = .59)	Maximum	<i>M</i>	<i>SD</i>	Skewness (<i>SE</i> = .33)	Kurtosis (<i>SE</i> = .66)
MAT	58	30.48	12.14	0.28	-0.11	43	24.51	8.55	-0.18	0.02
L1 RAN G1	228	70.61	43.57	2.13	4.33	613	111.10	109.80	3.10	10.36
L1 PA G1	20	9.58	6.59	0.13	-1.67	18	6.53	3.63	1.58	3.27
L1 WRA G1	43	8.48	6.65	2.58	10.84	49	14.63	9.05	1.62	3.94
L1 WRA G2	51	10.05	8.77	2.80	9.09	48	23.57	10.97	0.16	-0.46
L1 WRF G1	20	6.83	5.24	1.04	0.14	43	9.39	8.29	2.04	5.76
L1 WRF G2	28	8.66	6.09	1.36	1.57	55	18.47	10.38	1.17	2.26
E RAN G1	136	51.2	19.26	2.24	6.43	168	55.02	20.05	3.66	20.03
E PA G1	36	21.06	8.41	-0.72	0.11	33	23.08	6.37	-1.05	2.02
E WRA G1	75	43.08	16.9	-0.50	-0.24	93	32.78	18.48	0.72	1.37
E WRA G2	78	55.06	13.97	-0.84	0.33	81	52.37	12.44	-0.91	5.45
E WRF G1	71	41.25	18.88	-0.29	-1.08	80	26.73	15.1	1.04	2.22
E WRF G2	77	54.97	14.73	-1.04	0.70	88	49.37	15.04	-0.40	1.32

Note. MAT = Matrix Analogies Reasoning; RAN = Rapid Automatized Naming; PA = phonological awareness; WRA = word reading accuracy; WRF = word reading fluency; L1 = first language; E = English; G1 = Grade 1; G2 = Grade 2.

Regression and Commonality Analyses

Analytic strategy. Based on the key research questions and the results of the correlational analyses, hierarchical regression analyses were conducted to identify significant cross-language relationships for word reading accuracy and fluency in the Chinese–English and Spanish–English bilinguals, beyond within-language control variables. We first carried out four regression analyses on the combined sample. Grade 2 word reading accuracy or fluency in L1 or English acted as the dependent variable, respectively. For each regression analysis, language group (Spanish or Chinese) was entered in Step 1, followed by the main effect and interaction of the different independent variables and language group. Specifically, nonverbal reasoning and the interaction of this variable with language group were entered in Step 2. Within-language phonological awareness and the interaction of this variable with language group were entered in Step 3. Within-language

RAN and the interaction of RAN with language group were entered in Step 4. Step 5 included the autoregressor effect of Grade 1 word reading accuracy or fluency (congruent with the dependent variable) on Grade 2 word reading and the interaction of the autoregressor by language group. In Step 6, cross-language word reading accuracy or fluency in Grade 2 (congruent with the dependent variable) was added to test cross-language transfer of the construct, together with the interaction of language group and the cross-language variable.

Notably, our cross-language relationships were concurrent as we used Grade 2 cross-language variables to predict Grade 2 reading outcomes. This decision was made because the children's reading skills were more developed in Grade 2 and, consequently, the reading measures administered in Grade 2 captured more variance than the reading measures administered in Grade 1. To save space, the regression tables for the joint

Table 2

Pearson Correlations Among All Variables for the Chinese–English Bilinguals (Above the Diagonal) and the Spanish–English Bilinguals (Below the Diagonal)

Variable	1	2	3	4	5	6	7	8	9	10	11	12	13
1. MAT	—	-.03	.33**	.27*	-.01	.18	.10	-.35**	.43***	.48***	.44***	.41***	.36**
2. L1 RAN G1	-.04	—	.08	-.33**	-.28*	-.16	-.31*	.17	-.03	-.10	-.05	-.17	-.22
3. L1 PA G1	.13	-.27	—	-.02	-.11	.10	.04	-.25*	.64***	.50***	.42***	.44***	.31*
4. L1 WRA G1	.07	-.38**	.55**	—	.67***	.66***	.60***	-.27*	.13	.34**	.32*	.35**	.38**
5. L1 WRA G2	.01	-.44***	.49***	.65***	—	.57***	.83***	-.24	-.03	.25	.19	.23	.31*
6. L1 WRF G1	.17	-.28*	.68***	.68***	.50***	—	.77***	-.27*	.29*	.37**	.45***	.34**	.41***
7. L1 WRF G2	.09	-.28*	.60***	.70***	.78***	.77***	—	-.29*	.18	.39***	.37**	.38**	.45***
8. E RAN G1	-.26	.22	-.47**	-.28*	-.34*	-.47***	—	-.45**	-.57***	-.54***	-.61***	-.65***	—
9. E PA G1	.24	-.09	.45***	.08	.21	.26	.27	-.68**	—	.77***	.81***	.72***	.71***
10. E WRA G1	.21	-.13	.65***	.31*	.38**	.71***	.57***	-.59***	.45***	—	.90***	.94***	.88***
11. E WRA G2	.31*	-.11	.62***	.30*	.46***	.51***	.64***	-.76***	.58***	.65***	—	.87***	.91***
12. E WRF G1	.12	-.11	.61***	.35*	.30*	.73***	.56***	-.58***	.39***	.92***	.64***	—	.89***
13. E WRF G2	.21	-.06	.49***	.32*	.51***	.59***	.72***	.68***	.45***	.61***	.79***	.66***	—

Note. MAT = Matrix Analogies Reasoning; RAN = Rapid Automatized Naming; PA = phonological awareness; WRA = word reading accuracy; WRF = word reading fluency; L1 = first language; E = English; G1 = Grade 1; G2 = Grade 2.

p* < .05. *p* < .01. ****p* < .001.

analysis are not included in the article but are available in Tables S1-S4 of the online supplemental materials.

Considering most interaction terms were significant in the regression models described above, separate regressions were then conducted for the two groups. In each separate regression analysis, nonverbal reasoning, within-language phonological awareness, and within-language RAN were entered in the first three steps. Grade 1 word reading accuracy or fluency (congruent with the dependent variable) was added in Step 4 to control for the autoregressor effect of Grade 1 reading on Grade 2 reading. Finally, in Step 5, cross-language word reading accuracy or fluency in Grade 2 (congruent with the dependent variable) was added to test cross-language transfer of the construct.

To gain a better understanding of the cross-language relationships, we conducted commonality analyses separately for the Chinese-English and Spanish-English bilinguals to further clarify the similarities and differences between groups (Pedhazur, 1997). The motivation for conducting a commonality analysis was to understand the common and unique contributions of the predictor variables to the outcome variables. It is beneficial to supplement a regression analysis with a commonality analysis because it addresses collinearity in the data. Considering that the variables under study were often highly correlated with one another, a commonality analysis can establish the relative importance of these variables while accounting for the relationships among the independent variables (Mood, 1971; Newton & Spurrell, 1967; Nimon & Reio, 2011; Pedhazur, 1997; Zientek & Thompson, 2010). Additionally, a commonality analysis produces both beta weights and structural coefficients. Structural coefficients are Pearson correlation coefficients between predictor variables and the predicted outcome. It is beneficial to examine both beta weights and structural coefficients, especially when multicollinearity may occur, as the influence of a predictor variable on an outcome variable may be shared with another variable's beta weight (Courville & Thompson, 2001; Nimon & Reio, 2011; Thompson, 2006; Zientek & Thompson, 2010). To facilitate the reporting of the results, we present the variance unique to each predictor and the sum of all the variance components containing the cross-language predictor. The complete commonality analysis for each dependent variable and each group is available in Tables S5-S6 of the online supplemental materials.

The regression and commonality analyses are summarized in Tables 3 (regression), 4 (regression), 5 (commonality), 6 (regres-

sion), 7 (regression), and 8 (commonality). All regressions, whether conducted jointly or separately for the two groups, satisfied the assumptions of normality, homogeneity, and independence. Since the beta weights produced by the regression analyses and the commonality analyses were identical, they are presented only in the regression tables to avoid redundancy.

Cross-language transfer of word reading accuracy. The left panel of Table 3 displays the results of the regression analysis predicting Grade 2 English word reading accuracy for the Chinese speakers. In this regression analysis, Grade 1 English phonological awareness and Grade 1 English word reading accuracy both explained unique variance. The right panel of Table 3 presents the regression analysis with Grade 2 English word reading accuracy as the dependent variable for the Spanish speakers. Grade 1 English RAN and Grade 1 English word reading accuracy were unique predictors of Grade 2 English word reading accuracy for Spanish-English bilinguals. Cross-language relationships were not significant for either group when predicting English word reading accuracy. Specifically, when entered in the last step, Grade 2 Chinese word reading accuracy did not explain unique variance in Grade 2 English word reading accuracy. Similarly, Grade 2 Spanish word reading accuracy did not predict unique variance in Grade 2 English word reading accuracy when entered last.

The regression analysis predicting Grade 2 Chinese word reading accuracy for Chinese-English bilinguals is displayed in the left panel of Table 4. Only Grade 1 Chinese word reading accuracy was a unique predictor of Grade 2 Chinese word reading accuracy. Grade 2 English word reading accuracy did not contribute any unique variance to Grade 2 Chinese word reading accuracy. Thus, there was no cross-language transfer of word reading accuracy between Chinese and English. The right panel of Table 4 presents the results of the regression analysis examining Spanish word reading accuracy for the Spanish-English bilinguals. Among the three Grade 1 Spanish measures, only the autoregressor measure of Spanish word reading accuracy uniquely predicted the dependent variable. Importantly, Grade 2 English word reading accuracy was a unique predictor of Grade 2 Spanish word reading accuracy above and beyond all within-language controls. Thus, in the Spanish-English bilinguals, there was a significant cross-language effect of word reading accuracy from English to Spanish. Significant transfer of word reading accuracy for the Spanish-English bilinguals, but not the Chinese-English bilinguals, was confirmed by a significant interaction of English word reading accuracy with

Table 3
Regression Models Predicting Grade 2 English Word Reading Accuracy From Grade 2 L1 Word Reading Accuracy

Step	Variable	Chinese-English bilinguals			Spanish-English bilinguals		
		Final <i>B</i>	SE (<i>B</i>)	β	Final <i>B</i>	SE (<i>B</i>)	β
1	MAT	-.02	.07	-.01	.17	.13	.12
2	E PA G1	.51	.14	.31***	.19	.23	.10
3	E RAN G1	-.02	.05	-.02	-.28	.08	-.45***
4	E WRA G1	.54	.08	.65***	.18	.07	.26*
5	L1 WRA G2	.06	.09	.03	.19	.11	.17

Note. MAT = Matrix Analogies Reasoning; RAN = Rapid Automatized Naming; PA = phonological awareness; WRA = word reading accuracy; L1 = first language; E = English; G1 = Grade 1; G2 = Grade 2.
* $p < .05$. *** $p < .001$.

Table 4
Regression Models Predicting Grade 2 L1 Word Reading Accuracy From Grade 2 English Word Reading Accuracy

Step	Variable	Chinese-English bilinguals			Spanish-English bilinguals		
		Final <i>B</i>	SE (<i>B</i>)	β	Final <i>B</i>	SE (<i>B</i>)	β
1	MAT	-.15	.08	-.21	-.19	.13	-.15
2	L1 PA G1	-.08	.14	-.06	-.16	.44	-.05
3	L1 RAN G1	-.01	.02	-.06	-.02	.01	-.23
4	L1 WRA G1	.89	.14	.67***	.61	.15	.50***
5	E WRA G2	.06	.07	.09	.32	.12	.36**

Note. MAT = Matrix Analogies Reasoning; RAN = Rapid Automatized Naming; PA = phonological awareness; WRA = word reading accuracy; L1 = first language; E = English; G1 = Grade 1; G2 = Grade 2.
** $p < .01$. *** $p < .001$.

language group when L1 word reading accuracy was the dependent variable in the joint analysis.

The left panel of Table 5 presents the structural coefficients, proportions of unique and common variance and the corresponding percentages of explained variance for the model predicting English word reading accuracy. As explained earlier, we only present the sum of all the variance components, including the cross-language predictor. When English word reading accuracy was the dependent variable, structural coefficients of L1 word reading accuracy were weak (.21) for the Chinese-English bilinguals and moderate (.56) for the Spanish-English bilinguals. Virtually no unique variance was explained by L1 word reading accuracy for both groups (0.11% for the Chinese-English bilinguals and 3.26% for the Spanish-English bilinguals). Additionally, L1 word reading accuracy explained 4.22% of the common variance in English word reading accuracy for the Chinese-English bilinguals and 27.72% of the common variance for the Spanish-English bilinguals. The right panel of Table 5 displays the results predicting L1 word reading accuracy. Structural coefficients of English word reading accuracy were weak (.27) for the Chinese-English bilinguals and

moderate (.61) for the Spanish-English bilinguals. English word reading accuracy explained 1.11% of the unique variance for the Chinese-English bilinguals and 12.74% of the unique variance for the Spanish-English bilinguals. Relatedly, English word reading accuracy accounted for a substantial portion of common variance, with the other predictors, for the Spanish-English bilinguals (24.42%), but not for the Chinese-English bilinguals (6.48%).

Cross-language transfer of word reading fluency. The left panel of Table 6 displays the results of the regression analysis predicting Grade 2 English word reading fluency for the Chinese-English bilinguals. English phonological awareness, English RAN, and English word reading fluency in Grade 1 all contributed unique variance to Grade 2 English word reading fluency. Notably, Grade 2 Chinese word reading fluency, entered in the last step, was also a significant predictor of Grade 2 English word reading fluency. The right panel of Table 6 presents the regression model predicting Grade 2 English word reading fluency for the Spanish-English bilinguals. Grade 1 English word reading fluency was the only unique predictor among the three within-language variables. Importantly, Grade 2 Spanish word reading fluency contributed

Table 5
Commonality Analysis for Predicting Grade 2 English and L1 Word Reading Accuracy

Unique to	E WRA G2						L1 WRA G2					
	Chinese-English bilinguals			Spanish-English Bilinguals			Chinese-English bilinguals			Spanish-English bilinguals		
	Structural coefficient	Variance coefficient	R^2									
MAT	.47	.000	0.02%	.38	.012	1.80%	-.01	.032	6.51%	.00	.019	3.42%
E PA G1	.88	.034	3.93%	.71	.005	0.74%						
E RAN G1	-.58	.000	0.04%	-.92	.082	12.10%						
E WRA G1	.98	.125	14.60%	.80	.042	6.20%						
L1 WRA G2	.21	.001	0.11%	.56	.022	3.26%						
L1 PA G1							-.16	.003	0.60%	.65	.001	0.23%
L1 RAN G1							-.40	.003	0.67%	-.58	.043	7.66%
L1 WRA G1							.95	.344	69.63%	.87	.160	28.20%
E WRA G2							.27	.006	1.11%	.61	.072	12.74%
Common variance ^a	.036	4.22%		.188	27.72%		.032	6.48%		.138	24.42%	
Total ^b	.853			.678			.494			.565		

Note. MAT = Matrix Analogies Reasoning; RAN = Rapid Automatized Naming; PA = phonological awareness; WRA = word reading accuracy; L1 = first language; E = English; G1 = Grade 1; G2 = Grade 2.

^a Common variance of E WRA G2 columns represents the sum of all the variance components containing L1 WRA G2. Common variance of L1 WRA G2 columns represents the sum of all the variance components containing E WRA G2. Common variance of other components is not included here to save space. ^b Total represents the total amount of unique and shared variance of all the variables in the model.

Table 6
Regressions Predicting Grade 2 English Word Reading Fluency From Grade 2 L1 Word Reading Fluency

Step	Variable	Chinese-English bilinguals			Spanish-English bilinguals		
		Final <i>B</i>	<i>SE</i> (<i>B</i>)	β	Final <i>B</i>	<i>SE</i> (<i>B</i>)	β
1	MAT	-.06	.07	-.05	.11	.15	.06
2	E PA G1	.32	.14	.18*	.07	.27	.03
3	E RAN G1	-.13	.05	-.17*	-.24	.10	-.31
4	E WRF G1	.49	.07	.62***	.21	.11	.21*
5	L1 WRF G2	.34	.14	.14*	.64	.15	.44***

Note. MAT = Matrix Analogies Reasoning; RAN = Rapid Automatized Naming; PA = phonological awareness; WRF = word reading fluency; L1 = first language; E = English; G1 = Grade 1; G2 = Grade 2.
 * $p < .05$. *** $p < .001$.

unique variance to Grade 2 English word reading fluency beyond all within-language controls.

The regression model predicting Grade 2 Chinese word reading fluency for the Chinese-English bilinguals is presented in the left panel of Table 7. Both Grade 1 Chinese RAN and Grade 1 Chinese word reading fluency were unique predictors. Grade 2 English word reading fluency was a marginally significant predictor ($p = .068$). The regression analysis with Grade 2 Spanish word reading fluency as the dependent variable is displayed in the right panel of Table 7. Among the three Grade 1 Spanish variables, only Grade 1 word reading fluency accounted for unique variance in Grade 2 Spanish word reading fluency. It is noteworthy that Grade 2 English word reading fluency was also a unique predictor of Grade 2 Spanish word reading fluency. The results from the joint analysis confirmed the cross-language results. No significant interaction was found when English reading fluency was the dependent variable, suggesting that patterns of transfer were not statistically different across groups. When predicting L1 word reading fluency, there was a significant interaction of the cross-language predictor with language group. Overall, the results of reading fluency across languages suggested that for the Chinese-English bilinguals, there was a significant cross-language effect from Chinese to English and a marginally significant effect from English to Chinese. For the Spanish group, there was bidirectional transfer of word reading fluency between Spanish and English.

The left panel of Table 8 presents the structural coefficients, the proportions of unique and common variance, and the corresponding percentages of explained variance for the model predicting

English word reading fluency. The structural coefficients between L1 word reading fluency and English word reading fluency were moderate (.49) to strong (.86) for the Chinese-English bilinguals and Spanish-English bilinguals, respectively. L1 word reading fluency explained 1.91% of the unique variance for the Chinese-English bilinguals and 18.16% for the Spanish-English bilinguals. Notably, despite the difference in the amount of variance explained, regression analyses showed that L1 reading fluency was a significant unique predictor of English word reading fluency for both groups. In addition, L1 word reading fluency explained about 22.53% of the common variance in the Chinese-English bilinguals and 56.34% of the common variance in the Spanish-English bilinguals. Finally, as shown in the right panel of Table 8, when L1 word reading fluency was the dependent variable, structural coefficients were moderate (.56) to strong (.85) for the Chinese-English bilinguals and Spanish-English bilinguals, respectively. English word reading fluency uniquely accounted for 3.21% and 15.75% of explained variance for the Chinese-English bilinguals and Spanish-English bilinguals, respectively. Additionally, English word reading fluency accounted for 27.52% and 55.66% of the variance jointly with other variables for the two groups, respectively.

Discussion

The present study examined cross-language transfer of word reading accuracy and fluency. The participants of the study were Spanish-English and Chinese-English bilinguals, whose first lan-

Table 7
Regressions Predicting Grade 2 L1 Word Reading Fluency From Grade 2 English Word Reading Fluency

Step	Variable	Chinese-English bilinguals			Spanish-English bilinguals		
		Final <i>B</i>	<i>SE</i> (<i>B</i>)	β	Final <i>B</i>	<i>SE</i> (<i>B</i>)	β
1	MAT	-.04	.04	-.07	-.11	.10	-.09
2	L1 PA G1	-.04	.08	-.05	.17	.31	.06
3	L1 RAN G1	-.02	.01	-.16*	-.01	.01	-.12
4	L1 WRF G1	.81	.10	.69***	.57	.15	.45***
5	E WRF G2	.07	.04	.17*	.30	.07	.43***

Note. MAT = Matrix Analogies Reasoning; RAN = Rapid Automatized Naming; PA = phonological awareness; WRF = word reading fluency; L1 = first language; E = English; G1 = Grade 1; G2 = Grade 2.
 * $p = .068$. ** $p < .05$. *** $p < .001$.

Table 8

Commonality Analysis for Predicting Grade 2 English and L1 Word Reading Fluency

Unique to	E WRF G2						L1 WRF G2					
	Chinese–English bilinguals			Spanish–English bilinguals			Chinese–English bilinguals			Spanish–English bilinguals		
	Structural coefficient	Variance coefficient	R^2 (%)	Structural coefficient	Variance coefficient	R^2 (%)	Structural coefficient	Variance coefficient	R^2 (%)	Structural coefficient	Variance coefficient	R^2 (%)
MAT	.39	.002	0.22%	.26	.004	0.50%	.13	.004	0.62%	.10	.008	1.10%
E PA G1	.78	.015	1.77%	.55	.001	0.07%						
E RAN G1	-.71	.018	2.16%	-.81	.039	5.54%						
E WRF G1	.97	.132	15.70%	.79	.025	3.66%						
L1 WRF G2	.49	.016	1.91%	.86	.126	18.20%						
L1 PA G1							.05	.002	0.26%	.71	.002	0.23%
L1 RAN G1							-.38	.024	3.60%	-.33	.013	1.79%
L1 WRF G1							.95	.396	60.63%	.91	.090	12.42%
E WRF G2							.56	.021	3.21%	.85	.114	15.75%
Common variance ^a	.189	22.53%		.391	56.34%		.180	27.52%		.403	55.66%	
Total ^b	.839			.694			.654			.724		

Note. MAT = Matrix Analogies Reasoning; RAN = Rapid Automatized Naming; PA = phonological awareness; WRF = word reading fluency; L1 = first language; E = English; G1 = Grade 1; G2 = Grade 2.

^a Common variance of E WRF G2 columns represents the sum of all the variance components containing L1 WRF G2. Common variance of L1 WRF G2 columns represents the sum of all the variance components containing E WRF G2. Common variance of other components is not included here to save space. ^b Total represents the total amount of unique and shared variance of all the variables in the model.

guages differ with respect to how the orthography represents the oral language. By comparing the transfer patterns between the two groups of children, we sought to understand script-specific and script-universal processes in cross-language relations for word reading accuracy and fluency. With respect to word reading accuracy, we predicted that there would be transfer between Spanish and English due to the shared script and the overlap in grapheme-to-phoneme correspondences and that there would not be transfer between Chinese and English due to script differences. Regarding word reading fluency, we predicted that transfer would occur for both Spanish–English and Chinese–English bilinguals as the ability to synchronize and integrate information from phonological, morphological, and orthographic systems should be related across the L1 and L2, regardless of the degree of overlap between the two languages.

As expected, we found cross-language relations for word reading accuracy in Spanish–English bilinguals. Grade 2 English word reading accuracy predicted unique variance in Grade 2 Spanish word reading accuracy after controlling for nonverbal reasoning, Grade 1 Spanish phonological awareness, Grade 1 RAN, and Grade 1 Spanish word reading accuracy. Furthermore, commonality analysis showed that Grade 2 English word reading accuracy also explained a large amount of variance in Grade 2 Spanish word reading accuracy jointly with other predictors. Our results add to a growing body of research demonstrating cross-language transfer of phonological and word reading skills in Spanish–English bilinguals (August & Shanahan, 2006; Dressler & Kamil, 2006; Gholamain & Geva 1999; Gottardo, 2002; Lindsey et al., 2003; Manis et al., 2004; Páez & Rinaldi, 2006). Since Spanish and English are both represented by alphabetic scripts, they require the same component skills in reading acquisition. Phonological awareness, letter knowledge, and decoding strategies are important for learning to read both languages. The association observed between English and Spanish word reading accuracy suggests that for bilinguals, structural similarities in L1 and L2 facilitate cross-language relationships for word reading accuracy.

Interestingly, the crossover effect of word reading accuracy in Spanish–English bilinguals was only found from English to Spanish, not in the other direction. Grade 2 Spanish word reading accuracy was not a unique predictor of Grade 2 English word reading accuracy, although there was a strong structural coefficient and a large amount of shared variance with the other predictors in the model. The unidirectional association was likely related to the fact that the children received their formal reading instruction in English and, as a result, their English reading skills were more advanced than their Spanish reading skills. Extant research has demonstrated similar patterns of transfer where language and literacy skills in the dominant language facilitate acquisition of literacy skills in the less proficient language (Deacon, Wade-Woolley & Kirby, 2007; Pasquarella, Chen, Lam, Luo, & Ramirez, 2011; Zhang et al., 2010). Alternately, as suggested by a reviewer, the direction of the cross-language association may derive from contrasting patterns of regularity in the two languages. Grapheme-to-phoneme correspondences are far less regular and therefore more difficult to acquire in English than in Spanish. Therefore, English may provide a better foundation for transfer than Spanish as knowledge of Spanish grapheme-to-phoneme correspondences may be too regular to add any value to English reading accuracy. Additional research is needed to determine whether these explanations are complementary or whether one will prevail. Our findings, combined with those of previous research, suggest that possible factors that determine the direction of cross-language relations of literacy skills include bilingual students' relative proficiency in their L1 and L2 and the typological characteristics of the scripts involved.

Our study did not find significant relationships between L1 and L2 word reading accuracy skills for the Chinese–English bilinguals. The cross-language structural coefficients were weak, and the amount of variance shared with other predictors was also small. These findings were consistent with the research conducted by Gottardo et al. (2001) and Wang et al. (2005). Lack of associations between Chinese and English reinforces the notion that transfer of

word reading accuracy is a script-specific process that operates on similarities between L1 and L2. In addition, there is evidence that contextual factors, such as the language learning environment and the instructional method, play a role in transfer of reading skills across languages. Similar to Gottardo et al. and Wang et al., our study was conducted in the North American context. Research that reported a significant correlation between Chinese and English word reading was conducted by Keung and Ho (2009) in Hong Kong. North American children are likely to use phonological strategies to read English because phonics instruction is an important component in early reading programs. Children in Hong Kong, on the other hand, may be taught using the “look-and-say” method and therefore may be more inclined to use Chinese strategies, such as memorizing the whole word, to read English words (Keung & Ho, 2009; Shu & Anderson, 1999). It is possible that overlap in instructional methods and reading strategies between Chinese and English strengthens the cross-language connection of word reading skills in children in Hong Kong. The effect of educational context on cross-language transfer needs to be systematically investigated by future studies that include comparable cross-cultural samples.

With respect to word reading fluency, bidirectional relationships were found for the Spanish–English bilinguals. Grade 2 Spanish word reading fluency was a unique predictor of Grade 2 English word reading fluency after controlling for nonverbal reasoning and Grade 1 English phonological awareness, Grade 1 English RAN, and Grade 1 English word reading fluency. Similarly, Grade 2 English word reading fluency explained unique variance in Grade 2 Spanish word reading fluency after controlling for nonverbal reasoning and the Grade 1 Spanish measures. Additionally, commonality analysis showed that the cross-language structural coefficients were strong, and the cross-language predictor explained a very large amount of variance in word reading fluency jointly with other predictors in both models. Thus, our results, obtained with stringent within-language controls, confirm and extend the work of De Ramírez and Shapiro (2007), who reported significant correlations between Spanish and English text reading fluency. Therefore, word reading fluency is related for Spanish and English, two alphabetic orthographies, despite the fact that Spanish has more transparent grapheme-to-phoneme correspondences than English.

Importantly, cross-language transfer of word reading fluency was also observed in the Chinese–English bilinguals, whose two languages are represented by typologically different orthographies. For this group, Grade 2 Chinese word reading fluency predicted unique variance in Grade 2 English word reading fluency after accounting for nonverbal reasoning and Grade 1 English control variables. The prediction from Grade 2 English word reading fluency to Grade 2 Chinese word reading fluency was marginally significant and displayed a moderate cross-language structural coefficient. It is noteworthy that the cross-language predictor explained a large amount of variance in word reading fluency together with the other predictors in both models. By contrast, there was no unique prediction, weak cross-language structural coefficients, and only a small amount of shared variance in the models predicting word reading accuracy for the Chinese–English bilinguals.

Since cross-language relationships occurred for word reading fluency in both Spanish–English and Chinese–English bilinguals, it appears that the mechanism underlying word reading fluency is

largely script universal and not heavily influenced by differences in bilinguals’ L1 and L2. Word reading is based on perceptual, phonological, orthographic, and morphological processes, and reading fluency is determined by both the speed of processing within each system and synchronization and integration of the different systems (Breznitz, 2003, 2006; Breznitz & Berman, 2003; Seidenberg, 2005). Although the different systems function in somewhat different ways in Spanish, English, and Chinese, the synchronization and integration process may be a universal aspect of reading across the three languages. Our findings suggest that having well-developed reading fluency in one orthography contributes to reading fluency in another regardless of L1 and L2 differences.

On the other hand, there were differences in the patterns of relationships for word reading fluency in the Spanish–English and Chinese–English bilinguals. In some instances, the crossover effect was stronger in the Spanish–English bilinguals than the Chinese–English bilinguals, as indicated by the significant interaction of English word reading fluency and the children’s L1 (i.e., Spanish or Chinese) when predicting L1 word reading fluency. Commonality analyses also demonstrated stronger cross-language structural coefficients and more unique and shared variance between Spanish and English than between Chinese and English when predicting reading fluency. This result is based on the fact that accurate reading of words is necessary for word reading fluency. Our sample consisted of young children some of whom were fluent readers while others had not reached automaticity in word recognition. This variability in the sample resulted in an overlap between word reading accuracy and fluency skills, which was expected to be higher for the Spanish speakers than for the Chinese speakers. In other words, for Spanish–English bilinguals, cross-language relations for word reading fluency might be strengthened by cross-language relations for word reading accuracy. This facilitation does not occur for Chinese–English bilinguals, as word reading accuracy is not associated across the two languages.

The results of the present study contribute to the theory of cross-language transfer. We outlined two different perspectives toward cross-language transfer at the beginning of the article. According to the script-dependent perspective, cross-language transfer of a reading skill is based upon similarities across languages—the greater the overlap, the stronger the association of the skills across languages. This type of transfer is unlikely to occur if two languages do not share common features (Bialystok, Majumder, & Martin, 2003; Koda, 2007; Osgood, 1949; Pasquarella et al., 2011; Ziegler & Goswami, 2005). Alternatively, the script-universal perspective conceptualizes cross-language transfer as operating under cognitive and linguistic processes that are important for reading in all languages (Geva & Siegel, 2000). For bilingual children, a script-universal process transfers in the sense that it is recruited when reading both L1 and L2. This type of transfer is not conditioned by overlapping structural features.

Our results provide strong evidence that both types of cross-language transfer occur in bilingual children’s reading development. While transfer of word reading accuracy hinges upon shared structures between L1 and L2, transfer of word reading fluency occurs regardless of structural similarities. In addition, our results point to the possibility that both script-specific and script-universal processes underlie the transfer of a single construct due to its

complex and multifaceted nature. As previously mentioned, although word reading fluency is largely a script-universal process, it involves a script-specific component for less proficient readers because they apply decoding skills to read novel words encountered in the fluency test. Our study represents the first step toward disentangling the two types of transfer in bilingual children's reading development. Future research should examine the nature of transfer for other reading skills, such as morphological awareness, vocabulary, and reading comprehension.

The findings of our study have practical implications for the assessment and classification of bilingual children. Due to the well-known finding that phonological awareness is highly related between bilingual children's L1 and L2 regardless of the typological distance between the two languages, it is commonly agreed that there is no need to delay assessment for bilingual children until they develop sufficient oral proficiency in L2 (Cline & Frederickson, 1999; Geva & Herbert, 2012). Even for bilinguals from a nonalphabetic L1 background, assessment of phonological awareness can be effectively conducted in L1 for the purpose of predicting reading success in English. However, our findings indicate that unlike phonological awareness, word reading strategies are conditioned by script characteristics. Therefore, it cannot be assumed that older Chinese immigrant children will use alphabetic decoding strategies in reading English words. Instead, they may apply word reading strategies that are effective for reading Chinese, possibly leading to reading problems in English. Thus, explicit instruction in decoding strategies specific to English is necessary for immigrant children who are already literate in their L1, especially for those from a nonalphabetic background such as Chinese.

Another implication lies in using reading fluency as a potential cross-language screening measure for bilingual children. Although there is substantial evidence that word reading fluency is an effective screening measure for children who are native speakers of English (e.g., Compton et al., 2006, 2010; Fuchs et al., 2004), little is known about the value of using L1 reading fluency as a screening measure for bilingual children who are not yet literate in their L2. An important finding of the present study is that word reading fluency transfers between L1 and L2 in Spanish–English bilinguals as well as in Chinese–English bilinguals. In other words, L1 reading fluency can predict success in L2 reading fluency and is likely related to L2 reading comprehension for bilingual children from both alphabetic and nonalphabetic backgrounds (Kim, Wagner, & Lopez, 2012). Conversely, students with poor L1 fluency may be at risk of developing L2 reading problems regardless of their L1 backgrounds. Assessing reading fluency in the L1 may be especially useful for older immigrant children who are already literate in the L1 but are still building oral and reading vocabulary in the L2. Such assessment can avoid a further delay in identification of reading difficulties in this population.

The present study has several limitations. One limitation concerns measurement of the same reading and cognitive variables across Chinese, Spanish, and English. We used parallel standardized measures across languages when they were available. Because there were no Chinese standardized measures, we created our own experimental measures. All of our Chinese measures were highly reliable. They also demonstrated correlational patterns similar to those observed in previous research (e.g., Gottardo et al., 2001; McBride-Chang & Ho, 2005),

providing additional evidence of measurement validity. However, since our research did not focus on measurement issues, establishing validity of the Chinese measures through empirical methods goes beyond the scope of the present study. Future studies should develop multiple measures of these constructs and administer them to a larger number of participants to examine validity. Another limitation lies in the comparability of our Chinese–English bilinguals and Spanish–English bilinguals. Although the two groups of children were similar in many ways, including age, L1 exposure, and parental education, there were also some differences, especially in terms of enrollment in heritage language programs and attrition rates. Because these differences were also observed in previous studies comparing Chinese–English bilinguals and Spanish–English bilinguals (e.g., Bialystok et al., 2003; Chen, Ramirez, Luo, Geva, & Ku, 2012), we suspect that they represent systematic differences that exist between the two populations. Unfortunately, these were factors that we could not manipulate and attest to challenges in conducting research involving more than one group of bilingual children. Furthermore, our Chinese sample included both Mandarin- and Cantonese-speaking participants. Future research should replicate our findings with Chinese–English bilinguals who speak the same L1. Finally, future studies should be designed to specifically address factors that influence the direction and degree of cross-language transfer. For example, comparing the transfer patterns among balanced bilinguals, L1 dominant children, and L2 dominant children may shed light on how cross-language relations are influenced by language dominance.

To summarize, the present study provides novel insights into the underlying processes of transfer of word reading accuracy and word reading fluency in bilingual children. First, our results indicate that cross-language relations for word reading accuracy are conditioned by similarities between bilingual children's L1 and L2. Our study went beyond the previous research by including stringent control variables in the regressions and using commonality analyses to dissect unique and shared variance in cross-language relations. Second, our study reveals that transfer of word reading fluency is largely a script-universal process, similar to transfer of phonological awareness and working memory (Geva & Ryan, 1993; Geva & Siegel, 2000; Manis et al., 2004). Finally, our results suggest that cross-language relations are influenced by bilingual readers' relative proficiency in L1 and L2, as well as by contextual factors such as the language learning environment and instructional methods. Understanding transfer of word reading accuracy and word reading fluency and the factors that influence the direction and strength of transfer have important implications for the assessment and instruction of bilingual children.

References

Aaron, P. G., Joshi, M., & Williams, K. A. (1999). Not all reading disabilities are alike. *Journal of Learning Disabilities*, 32, 120–137. doi:10.1177/002221949903200203

Abu-Rabia, S., & Siegel, L. S. (2002). Reading, syntactic, orthographic, and working memory skills of bilingual Arabic-English speaking Canadian children. *Journal of Psycholinguistic Research*, 31, 661–678. doi:10.1023/A:1021221206119

August, D., & Shanahan, T. (Eds.). (2006). *Developing literacy in second-language learners: A report of the National Literacy Panel on Language-Minority Children and Youth*. Mahwah, NJ: Erlbaum.

Ball, E., & Blachman, B. A. (1991). Does phoneme awareness training in kindergarten make a difference in early word recognition and developmental spelling? *Reading Research Quarterly*, 26, 49–66. doi:10.1598/RRQ.26.1.3

Bialystok, E., Majumder, S., & Martin, M. M. (2003). Developing phonological awareness: Is there a bilingual advantage? *Applied Psycholinguistics*, 24, 27–44. doi:10.1017/S014271640300002X

Breznitz, Z. (2003). Speed of phonological and orthographic processing as factors in dyslexia: Electrophysiological evidence. *Genetic, Social, and General Psychology Monographs*, 129, 183–206.

Breznitz, Z. (2006). *Fluency in reading: Synchronization of processes*. Mahwah, NJ: Erlbaum.

Breznitz, Z., & Berman, L. (2003). The underlying factors of word reading rate. *Educational Psychology Review*, 15, 247–265. doi:10.1023/A:1024696101081

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, 41, 977–990. doi:10.3758/BRM.41.4.977

Cai, Q., & Brysbaert, M. (2010). SUBTLEX-CH: Chinese word and character frequencies based on film subtitles. *PLoS ONE*, 5(6), Article e10729. doi:10.1371/journal.pone.0010729

Chen, X., Ramirez, G., Luo, Y., Geva, E., & Ku, Y.-M. (2012). Comparing vocabulary development in Spanish- and Chinese-speaking ELLs: The effects of metalinguistic and sociocultural factors. *Reading and Writing*, 25, 1991–2020. doi:10.1007/s11145-011-9318-7

Clemens, N. H., Shapiro, E. S., & Thoemmes, F. (2011). Improving the efficacy of first grade reading screening: An investigation of word identification fluency with other early literacy indicators. *School Psychology Quarterly*, 26, 231–244. doi:10.1037/a0025173

Cline, T., & Frederickson, N. (1999). Identification and assessment of dyslexia in bi/multilingual children. *International Journal of Bilingual Education and Bilingualism*, 2, 81–93. doi:10.1080/13670059908667681

Compton, D. L., Fuchs, D., Fuchs, L. S., Bouton, B., Gilbert, J. K., Barquero, L. A., . . . Crouch, R. C. (2010). Selecting at-risk first-grade readers for early intervention: Eliminating false positives and exploring the promise of a two-stage gated screening process. *Journal of Educational Psychology*, 102, 327–340. doi:10.1037/a0018448

Compton, D. L., Fuchs, D., Fuchs, L. S., & Bryant, J. D. (2006). Selecting at-risk readers in first grade for early intervention: A two-year longitudinal study of decision rules and procedures. *Journal of Educational Psychology*, 98, 394–409. doi:10.1037/0022-0663.98.2.394

Courville, T., & Thompson, B. (2001). Use of structure coefficients in published multiple regression articles: β is not enough. *Educational and Psychological Measurement*, 61, 229–248.

Cuetos, F., Glez-Nosti, M., Barbón, A., & Brysbaert, M. (2011). SUBTLEX-ESP: Spanish word frequencies based on film subtitles. *Psicológica*, 32, 133–143.

Da Fontoura, H. A., & Siegel, L. S. (1995). Reading, syntactic, and working memory skills of bilingual Portuguese-English Canadian children. *Reading and Writing*, 7, 139–153. doi:10.1007/BF01026951

Deacon, S. H., Wade-Woolley, L., & Kirby, J. (2007). Crossover: The role of morphological awareness in French immersion children's reading. *Developmental Psychology*, 43, 732–746. doi:10.1037/0012-1649.43.3.732

Defior, S., Cary, L., & Martos, F. (2002). Differences in reading acquisition development in two shallow orthographies: Portuguese and Spanish. *Applied Psycholinguistics*, 23, 135–148. doi:10.1017/S0142716402000073

De Ramírez, R. D., & Shapiro, E. S. (2007). Cross-language relationship between Spanish and English oral reading fluency among Spanish-speaking English language learners in bilingual education classrooms. *Psychology in the Schools*, 44, 795–806. doi:10.1002/pits.20266

Dressler, C., & Kamil, M. L. (2006). First- and second-language literacy. In D. August & T. Shanahan (Eds.), *Developing literacy in second-language learners: A report of the National Literacy Panel on Language-Minority Children and Youth* (pp. 197–238). Mahwah, NJ: Erlbaum.

Durgunoğlu, A. Y. (2002). Cross-linguistic transfer in literacy development and implications for language learners. *Annals of Dyslexia*, 52, 189–204. doi:10.1007/s11881-002-0012-y

Durgunoğlu, A. Y., Nagy, W. E., & Hancin-Bhatt, B. J. (1993). Cross-language transfer of phonological awareness. *Journal of Educational Psychology*, 85, 453–465. doi:10.1037/0022-0663.85.3.453

Ehri, L. C. (1997). Sight word learning in normal readers and dyslexics. In B. Blachman (Ed.), *Foundations of reading acquisition and dyslexia* (pp. 163–186). Mahwah, NJ: Erlbaum.

Francis, D., Carlo, M., August, D., Kenyon, D., Malabonga, V., Caglarcan, S., & Louguit, M. (2001). *Test of Phonological Processing in Spanish*. Washington, DC: Center for Applied Linguistics.

Fuchs, L. S., Fuchs, D., & Compton, D. L. (2004). Monitoring early reading development in first grade: Word identification fluency versus nonsense word fluency. *Exceptional Children*, 71, 7–21.

Fuchs, L. S., Fuchs, D., Hosp, M. K., & Jenkins, J. R. (2001). Oral reading fluency as an indicator of reading competence: A theoretical, empirical and historical analysis. *Scientific Studies of Reading*, 5, 239–256. doi:10.1207/S1532799XSSR0503_3

Fuchs, L. S., Fuchs, D., & Maxwell, L. (1988). The validity of informal reading comprehension measures. *Remedial and Special Education*, 9, 20–28. doi:10.1177/074193258800900206

Genesee, F., & Geva, E. (2006). Cross-linguistic relationships in working memory, phonological processing, and oral language. In D. August & T. Shanahan (Eds.), *Developing literacy in second-language learners: A report of the National Literacy Panel on Language-Minority Children and Youth* (pp. 169–177). Mahwah, NJ: Erlbaum.

Genesee, F., Geva, E., Dressler, D., & Kamil, M. (2006). Synthesis: Cross-linguistic relationships. In D. August & T. Shanahan (Eds.), *Developing literacy in second-language learners: A report of the National Literacy Panel on Language-Minority Children and Youth* (pp. 153–174). Mahwah, NJ: Erlbaum.

Geva, E., & Farnia, F. (2012). Developmental changes in the nature of language proficiency and reading fluency paint a more complex view of reading comprehension in ELL and EL1. *Reading and Writing*, 25, 1819–1845. doi:10.1007/s11145-011-9333-8

Geva, E., & Herbert, K. (2012). Assessment and interventions in English language learners with LD. In B. Wong & D. Butler (Eds.), *Learning about learning disabilities* (4th ed., pp. 271–298). doi:10.1016/B978-0-12-388409-1.00010-2

Geva, E., & Ryan, E. B. (1993). Linguistic and cognitive correlates of academic skills in first and second languages. *Language Learning*, 43, 5–42. doi:10.1111/j.1467-1770.1993.tb00171.x

Geva, E., & Siegel, L. (2000). Orthographic and cognitive factors in the concurrent development of basic reading skills in two languages. *Reading and Writing*, 12, 1–30. doi:10.1023/A:1008017710115

Gholamain, M., & Geva, E. (1999). Orthographic and cognitive factors in the concurrent development of basic reading skills in English and Persian. *Language Learning*, 49, 183–217. doi:10.1111/0023-8333.00087

Good, R. H., Simmons, D. C., & Kame'enui, E. J. (2001). The importance and decision-making utility of a continuum of fluency-based indicators of foundational reading skills for third-grade high-stakes outcomes. *Scientific Studies of Reading*, 5, 257–288. doi:10.1207/S1532799XSSR0503_4

Gottardo, A. (2002). The relationship between language and reading skills in bilingual Spanish-English speakers. *Topics in Language Disorders*, 22, 46–70. doi:10.1097/00011363-200211000-00008

Gottardo, A., Yan, B., Siegel, L. S., & Wade-Woolley, L. (2001). Factors related to English reading performance in children with Chinese as a first language: More evidence for cross-language transfer of phonological processing. *Journal of Educational Psychology*, 93, 530–542. doi:10.1037/0022-0663.93.3.530

Gough, P. B., & Tunmer, W. E. (1986). Decoding, reading, and reading disability. *Remedial and Special Education*, 7(1), 6–10. doi:10.1177/074193258600700104

Hoover, W. A., & Gough, P. B. (1990). The simple view of reading. *Reading and Writing*, 2, 127–160. doi:10.1007/BF00401799

Jared, D., Cormier, P., Levy, B. A., & Wade-Woolley, L. (2011). Early predictors of biliteracy development in children in French immersion: A 4-year longitudinal study. *Journal of Educational Psychology*, 103, 119–139. doi:10.1037/a0021284

Jenkins, J. R., Fuchs, L. S., van den Broek, P., Espin, C., & Deno, S. L. (2003). Sources of individual differences in reading comprehension and reading fluency. *Journal of Educational Psychology*, 95, 719–729. doi:10.1037/0022-0663.95.4.719

Jímenez-González, J. E. (1997). A reading-level match study of phonemic processing underlying reading disability in a transparent orthography. *Reading and Writing*, 9, 23–40. doi:10.1023/A:1007925424563

Joshi, R. M., & Aaron, P. G. (2000). The component model of reading: Simple view of reading made a little more complex. *Reading Psychology*, 21, 85–97. doi:10.1080/02702710050084428

Kenny, D. A. (1975). Cross-lagged panel correlation: A test for spuriousness. *Psychological Bulletin*, 82, 887–903. doi:10.1037/0033-2909.82.6.887

Keung, Y. C., & Ho, C. S. H. (2009). Transfer of reading-related cognitive skills in learning to read Chinese (L1) and English (L2) among Chinese elementary school children. *Contemporary Educational Psychology*, 34, 103–112. doi:10.1016/j.cedpsych.2008.11.001

Kim, Y.-S., Wagner, R., & Lopez, D. (2012). Developmental relations between reading fluency and reading comprehension: Longitudinal study from Grade 1 to Grade 2. *Journal of Experimental Child Psychology*, 113, 93–111. doi:10.1016/j.jecp.2012.03.002

Koda, K. (2007). Reading and language learning: Crosslinguistic constraints on second language reading development. *Language Learning*, 57, 1–44. doi:10.1111/0023-8333.101997010-i1

Kuhn, M. R., Schwanenflugel, P. J., & Meisinger, E. B. (2010). Aligning theory and assessment of reading fluency: Automaticity, prosody and definitions of fluency. *Reading Research Quarterly*, 45, 230–251. doi:10.1598/RRQ.45.2.4

LaBerge, D., & Samuels, S. J. (1974). Toward a theory of automatic information processing in reading. *Cognitive Psychology*, 6, 293–323. doi:10.1016/0010-0285(74)90015-2

Leong, C. K., Cheng, P. W., & Mulcahy, R. (1987). Automatic processing of morphemic orthography by mature readers. *Language and Speech*, 30, 181–196.

Leong, C. K., Tse, S. K., Lon, K. Y., & Hau, K. T. (2008). Text comprehension in Chinese children: Relative contribution of verbal working memory, pseudoword reading, rapid automatized naming, and onset-rime phonological segmentation. *Journal of Educational Psychology*, 100, 135–149. doi:10.1037/0022-0663.100.1.135

Liao, C.-H., Georgiou, G. K., & Parrila, R. (2008). Rapid naming speed and Chinese character recognition. *Reading and Writing*, 21, 231–253. doi:10.1007/s11145-007-9071-0

Lindsey, K. A., Manis, F. R., & Bailey, C. E. (2003). Prediction of first-grade reading in Spanish-speaking English-language learners. *Journal of Educational Psychology*, 95, 482–494. doi:10.1037/0022-0663.95.3.482

Manis, F. R., Lindsey, K. A., & Bailey, C. E. (2004). Development of reading in Grades K–2 in Spanish-speaking English-language learners. *Learning Disabilities Research & Practice*, 19, 214–224. doi:10.1111/j.1540-5826.2004.00107.x

Marinova-Todd, S. H., Zhao, J., & Bernhardt, M. (2010). Phonological awareness skills in the two languages of Mandarin–English bilingual children. *Clinical Linguistics & Phonetics*, 24, 387–400. doi:10.3109/02699200903532508

Marston, D. (1989). A curriculum-based measurement approach to assessing academic performance: What is it and why do it. In M. R. Shinn (Ed.), *Curriculum-based measurement: Assessing special children* (pp. 18–78). New York, NY: Guilford Press.

Mattingly, I. G., & Xu, Y. (1993). Word superiority in Chinese. *Haskins Laboratories Status Report on Speech Research*, SR-113, 143–152.

McBride-Chang, C., Cho, J.-R., Liu, H., Wagner, R. K., Shu, H., Zhou, A., . . . Muse, A. (2005). Changing models across cultures: Associations of phonological awareness and morphological structure awareness with vocabulary and word recognition in second graders from Beijing, Hong Kong, Korea, and the United States. *Journal of Experimental Child Psychology*, 92, 140–160. doi:10.1016/j.jecp.2005.03.009

McBride-Chang, C., & Ho, C. S.-K. (2005). Predictors of beginning reading in Chinese and English: A 2-year longitudinal study of Chinese kindergartners. *Scientific Studies of Reading*, 9, 117–144. doi:10.1207/s1532799xssr0902_2

Mood, A. M. (1971). Partitioning variance in multiple regression analysis as a tool for developing learning models. *American Educational Research Journal*, 8, 191–202. doi:10.3102/00028312008002191

Morton, J. (1969). Interaction of information in word recognition. *Psychological Review*, 76, 165–178. doi:10.1037/h0027366

Naglieri, J. A. (1985). *Matrix Analogies Test—Expanded Form*. San Antonio, TX: Psychological Corporation.

National Institute of Child Health and Human Development. (2000). *Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction* (NIH Publication No. 00–4769). Washington, DC: U.S. Government Printing Office.

Newton, R. G., & Spurrell, D. J. (1967). A development of multiple regression for the analysis of routine data. *Applied Statistics*, 16, 51–64. doi:10.2307/2985237

Nimon, K., & Reio, T. G. (2011). Regression commonality analysis: A technique for quantitative theory building. *Human Resource Development Review*, 10, 329–340. doi:10.1177/1534484311411077

Odlin, T. (1989). *Language transfer: Cross-linguistic influence in language learning*. Cambridge, England: Cambridge University Press. doi:10.1017/CBO9781139524537

Osgood, C. E. (1949). The similarity paradox in human learning: A resolution. *Psychological Review*, 56, 132–143. doi:10.1037/h0057488

Páez, M., & Rinaldi, C. (2006). Predicting English word reading skills for Spanish-speaking students in first grade. *Topics in Language Disorders*, 26, 338–350. doi:10.1097/00011363-200610000-00006

Pasquarella, A., Chen, X., Lam, K., Luo, Y. C., & Ramirez, G. (2011). Cross-language transfer of morphological awareness in Chinese–English bilinguals. *Journal of Research in Reading*, 34, 23–42. doi:10.1111/j.1467-9817.2010.01484.x

Pedhazur, E. J. (1997). *Multiple regression in behavioral research: Explanation and prediction* (3rd ed.). Orlando, FL: Holt, Rinehart & Winston.

Perfetti, C. A. (1985). *Reading ability*. New York, NY: Oxford University Press.

Perfetti, C. A. (2003). The universal grammar of reading. *Scientific Studies of Reading*, 7, 3–24. doi:10.1207/S1532799XSSR0701_02

Perfetti, C. A., & Hogaboam, T. (1975). Relationship between single word decoding and reading comprehension skill. *Journal of Educational Psychology*, 67, 461–469. doi:10.1037/h0077013

Reicher, G. M. (1969). Perceptual recognition as a function of the meaningfulness of the stimulus material. *Journal of Experimental Psychology, 81*, 275–280. doi:10.1037/h0027768

Samuels, J. S. (2006). Toward a model of reading fluency. In J. S. Samuels & A. E. Farstrup (Eds.), *What research has to say about fluency instruction* (pp. 24–46). Newark, DE: International Reading Association.

Seidenberg, M. S. (2005). Connectionist model of word reading. *Current Directions in Psychological Science, 14*, 238–242. doi:10.1111/j.0963-7214.2005.00372.x

Seidenberg, M. S., & McClelland, J. L. (1989). A distributed developmental model of word recognition and naming. *Psychological Review, 96*, 523–568. doi:10.1037/0033-295X.96.4.523

Shaywitz, S. E., & Shaywitz, B. A. (2005). Dyslexia (specific reading disability). *Biological Psychiatry, 57*, 1301–1309. doi:10.1016/j.biopsych.2005.01.043

Shu, H., & Anderson, R. C. (1999). Learning to read Chinese: The development of metalinguistic awareness. In J. Wang, A. W. Inhoff, & H.-C. Chen (Eds.), *Reading Chinese script: A cognitive analysis* (pp. 1–18). Mahwah, NJ: Erlbaum.

Shu, H., Chen, X., Anderson, R. C., Wu, N., & Xuan, Y. (2003). Properties of school Chinese: Implications for learning to read. *Child Development, 74*, 27–47. doi:10.1111/1467-8624.00519

Stanovich, K. E. (1980). Toward an interactive-compensatory model of individual differences in development of reading fluency. *Reading Research Quarterly, 16*, 32–71. doi:10.2307/747348

Stanovich, K. E. (1988). Explaining the differences between the dyslexic and the garden-variety poor reader: The phonological-core variable-difference model. *Journal of Learning Disabilities, 21*, 590–604. doi:10.1177/002221948802101003

Stanovich, K. E. (1991). Word recognition: Changing perspectives. In R. Barr, M. L. Kamil, P. Mosenthal, & P. D. Pearson (Eds.) *Handbook of reading research* (Vol. 2, pp. 418–452). Mahwah, NJ: Erlbaum.

Stanovich, K. E. (1994). Romance and reality. *Reading Teacher, 47*, 280–291.

Swanson, H. L., & Berninger, V. (1995). The role of working memory in skilled and less skilled readers' comprehension. *Intelligence, 21*, 83–108. doi:10.1016/0160-2896(95)90040-3

Tan, L. H., & Perfetti, C. A. (1998). Phonological codes as early sources of constraint in Chinese word identification: A review of current discoveries and theoretical accounts. *Reading and Writing, 10*, 165–200. doi:10.1023/A:1008086231343

Thompson, B. (2006). *Foundations of behavioral statistics: An insight-based approach*. New York, NY: Guilford Press.

Tong, X., & McBride-Chang, C. (2010). Chinese-English biscriptal reading: Cognitive component skills across orthographies. *Reading and Writing, 23*, 293–310. doi:10.1007/s11145-009-9211-9

Torgesen, J. K., Wagner, R. K., & Rashotte, C. A. (1999). *TOWRE: Test of Word Reading Efficiency*. Austin, TX: PRO-ED.

Torgesen, J. K., Wagner, R. K., Rashotte, C. A., Rose, E., Lindamood, P., Conway, T., & Garvan, C. (1999). Preventing reading failure in young children with phonological processing disabilities: Group and individual responses to instruction. *Journal of Educational Psychology, 91*, 579–593. doi:10.1037/0022-0663.91.4.579

Venezky, R. L. (1970). *The structure of English orthography*. doi:10.1515/9783110804478

Wagner, R. K., Torgesen, J. K., & Rashotte, C. A. (1999). *CTOPP: Comprehensive Test of Phonological Processing*. Austin, TX: PRO-ED.

Wang, M., Perfetti, C. A., & Liu, Y. (2005). Chinese-English biliteracy acquisition: Cross-language and writing system transfer. *Cognition, 97*, 67–88. doi:10.1016/j.cognition.2004.10.001

Wimmer, H., & Goswami, U. (1994). The influence of orthographic consistency on reading development: Word recognition in English and German children. *Cognition, 51*, 91–103. doi:10.1016/0010-0277(94)90010-8

Woodcock, R. W. (1991). *Woodcock Reading Mastery Test—Revised*. Itasca, IL: Riverside.

Woodcock, R. W., & Munoz-Sandoval, A. (1995). *Woodcock Language Proficiency Battery—Revised, Spanish Form*. Itasca, IL: Riverside.

Yeung, P., Ho, C., Chik, P., Lo, L., Luan, H., Chan, D., & Ching, K. (2011). Reading and spelling Chinese among beginning readers: What skills make a difference? *Scientific Studies of Reading, 15*, 285–313. doi:10.1080/10888438.2010.482149

Zhang, J., Anderson, R. C., Li, H., Dong, Q., Wu, X., & Zhang, Y. (2010). Cross-language transfer of insight into the structure of compound words. *Reading and Writing, 23*, 311–336. doi:10.1007/s11145-009-9205-7

Ziegler, J. C., & Goswami, U. (2005). Reading acquisition, developmental dyslexia, and skilled reading across languages: A psycholinguistic grain size theory. *Psychological Bulletin, 131*, 3–29. doi:10.1037/0033-2909.131.1.3

Zientek, L. R., & Thompson, B. (2010). Using commonality analysis to quantify contributions that self-efficacy and motivational factors make in mathematics performance. *Research in the Schools, 17*(1), 1–11.

Received September 11, 2012

Revision received February 18, 2014

Accepted March 20, 2014 ■

Bilingual Phonological Awareness: Reexamining the Evidence for Relations Within and Across Languages

Lee Branum-Martin
Georgia State University

Sha Tao
Beijing Normal University

Sarah Garnaat
Alpert Medical School of Brown University

There is increasing interest in the role of phonological awareness across languages. Research is uncovering cross-language effects of phonological awareness upon English reading, even from nonalphabetic languages. However, little of this research has focused on examining the extent to which multiple measures of phonological awareness indicate a single construct within or across languages. This article updates 2 recent reviews of the literature by fitting rival a priori models of multiple measures in order to test within-language and across-language structure among multiple phonological awareness tasks. Although the number and types of languages covered were quite limited, the results demonstrate high cross-language correlations, suggesting that measurement error has attenuated prior estimates of the cross-language correlation of phonological tasks. The current results suggest that in alphabetic languages, there is empirical support for phonological awareness as a unitary ability within English and other languages. In Korean and in Spanish, phonological awareness may operate as a language-general construct. In Cantonese and Mandarin, the results were less clear. The results also highlight the limitations of the current research base and important areas for future investigation.

Keywords: bilingualism, phonological awareness, confirmatory factor analysis, structural equation model

There is growing interest in the cross-language role of phonological awareness, both as a construct by itself and because of its potentially important role in learning to read different languages (Branum-Martin, Tao, Garnaat, Bunta, & Francis, 2012; Durgunoğlu, Nagy, & Hancin-Bhatt, 1993; Melby-Lervåg & Lervåg, 2011). Phonological awareness, or the ability to recognize and manipulate linguistic sounds apart from their meanings, is a crucial skill in learning to read (National Institute of Child Health & Human Development, 2000; Rayner, Foorman, Perfetti, Pesetsky, & Seidenberg, 2001; Snow, Burns, & Griffin, 1998; Wagner & Torgesen, 1987). Diverse measures of phonological awareness have high but somewhat inconsistent correlations across languages, as shown in two recent meta-analyses

(Branum-Martin et al., 2012; Melby-Lervåg & Lervåg, 2011). However, method effects, such as using the same task in both languages (e.g., blending phonemes), could spuriously inflate cross-language correlations, leading researchers to believe that phonological awareness operates more similarly than it does in actuality. Relations across measures, both within and across languages, have implications for what we infer about phonological awareness: whether it is a single construct or multiple constructs, within language as well as across languages. In these prior analyses, such trait effects were not adequately distinguished from potential method effects.

We care about whether phonological awareness is a single construct because of its implications for instruction and intervention (Branum-Martin et al., 2006). If phonological awareness is a single construct across languages (i.e., is language general), phonological instruction should improve performance in both languages and potentially help reading, perhaps in either language. Alternatively, if phonological awareness is a language-specific ability—one construct in each language—instruction and interventions would have to be designed to the special cognitive and linguistic requirements of the separate abilities in those languages (as opposed to merely the specific sound structure and orthography of the particular language).

The issue of multiple measures of one or more constructs represents a classic multitrait, multimethod problem (Campbell & Fiske, 1959; Eid, Lischetzke, & Nussbeck, 2006). In this article, we reanalyze prior studies of cross-language phonological awareness to specifically test theoretical models of trait (construct) versus method effects.

This article was published Online First June 16, 2014.

Lee Branum-Martin, Department of Psychology, Georgia State University; Sha Tao, National Key Laboratory of Cognitive Neuroscience and Learning, Beijing Normal University; Sarah Garnaat, Department of Psychiatry and Human Behavior, Alpert Medical School of Brown University.

Sha Tao was partially supported by Natural Science Foundation of China Grant 30970908. The initial work for this article was done while the authors were at the Texas Institute for Measurement, Evaluation, and Statistics at the University of Houston. Lee Branum-Martin would like to thank David J. Francis for the initial suggestion that evolved into this study. He would also like to thank Jason L. Anthony, Ferenc Bunta, and Karla K. Stuebing for insightful comments on a draft of this article.

Correspondence concerning this article should be addressed to Lee Branum-Martin, Georgia State University Department of Psychology, P.O. Box 5010, Atlanta, GA 30302-5010. E-mail: BranumMartin@gsu.edu

Phonological Awareness Within and Across Languages

Although there is growing evidence of phonological awareness as a single ability across multiple tasks within many languages, such as Spanish, Dutch, Greek, Korean, and Chinese (see review by Branum-Martin et al., 2012), what facilitating or interfering role differing linguistic features may have when learning other languages is not clear. The role of phonology could be complex, involving multiple, language-specific constructs (abilities) in persons who speak more than one language (Grosjean & Li, 2013). Alternatively, phonology may involve but a single construct, which is domain general across languages (see the review by Thomas & van Heuven, 2005).

On a theoretical basis, it is not clear how consistent the effects of phonology in different languages should be and how those might facilitate or complicate our investigation of phonological awareness as a construct. Similarity in phonological, grammatical, and lexical features could ease performance of skills across languages (Flege, 1995), thereby suggesting that phonological abilities operate similarly, if not identically, across languages. Additionally, computational linguistic models of bilingualism imply that the overlap of orthography and phonology can be important in reading (Grosjean & Li, 2013; Thomas & van Heuven, 2005). However, these models suggest that similar letters or sounds can activate as well as inhibit linguistic recognition and production from one language to another. Thus, it is possible that cross-language influences in phonology could be helpful, be difficult to overcome, or have no effect.

The size of the linguistic units (e.g., syllables, onsets, rimes, phonemes) can also be important, especially when different languages emphasize different grain sizes (Frost, 1998; Grosjean, 2008; Grosjean & Li, 2013; Ziegler & Goswami, 2005, 2006). For example, alphabetic languages that emphasize phonemes, such as English or French, may foster different linguistic skills than do languages with more emphasis on larger grain sizes, such as Italian or Spanish (Ziegler & Goswami, 2005, 2006). Because the nature of phonological skills may differ across languages and in persons who speak more than one language, cross-language effects may not be simple.

Linguistic analysis and cognitive experiments suggest that there is a basis for expecting that, across languages, phonological abilities are likely activated by a number of reading tasks, even when phonological information is limited or the writing system is not alphabetic (Perfetti, 2003; Perfetti, Liu, & Tan, 2005; Perfetti & Zhang, 1995; Perfetti, Zhang, & Berent, 1992). It has been argued that "phonological processes occur as part of reading in all writing systems, with the details of the writing system influencing the details of phonological processing" (Perfetti & Zhang, 1995, p. 24). Writing systems represent spoken language, but they do not always do so reliably or consistently. Therefore, different writing systems will offer different degrees of facilitation and constraint in the representation of the sounds of speech (Flege, 1995; Frost, 1998; Grosjean, 2008; Grosjean & Li, 2013; Ziegler & Goswami, 2005, 2006). It is therefore plausible that phonological awareness is a single ability across languages, but the role of phonological awareness in learning to read may differ across writing systems.

We therefore have unclear and potentially conflicting theoretical expectations. Diverse measures of phonological awareness may indicate language-specific abilities (i.e., two or more factors that

may or may not be correlated), or these measures may indicate a single, domain-general ability across languages. Moreover, these structures of abilities could possibly differ in different languages (e.g., required skill structures may be different in English than they are in Chinese). Research must test these hypothetical possibilities while accounting for potential method effects that could be common in measures across languages.

Rationale for the Current Study

Two prior syntheses have examined cross-language correlations among phonological measures (Branum-Martin et al., 2012; Melby-Lervåg & Lervåg, 2011). The first cross-language meta-analysis of phonological awareness (Melby-Lervåg & Lervåg, 2011) examined the impact of instructional language (bilingual or only second language), home language (native or not reported), writing system (alphabetic or ideographic), and age upon 17 reported cross-language correlations between phonological awareness tasks among bilingual children. Among these studies, no significant effects were found for these four factors on the cross-language correlation between phonological awareness tasks.

A second meta-analysis substantially expanded upon these results to examine 101 correlations from 38 studies in nine languages (Branum-Martin et al., 2012). Cross-language correlations were found to differ strongly by the particular language but not necessarily by whether the writing system was alphabetic. There was a tendency for older samples of children to have lower cross-language correlations. Effects for linguistic features of the tasks were neither strong nor consistent (Branum-Martin et al., 2012). Together, these two studies suggest that cross-language relations among phonological awareness tasks are usually moderate to high: 0.39 to 0.86 (Branum-Martin et al., 2012; Melby-Lervåg & Lervåg, 2011).

The moderate to high correlations found in prior meta-analyses could simply reflect shared method variance between similar tasks: High cross-language correlation may be spurious. Although Branum-Martin et al. (2012) found no substantial difference in the cross-language correlation due to different types of task, only measures that were the same in each language were used.

The threat of method covariance across languages may obscure our attempts to understand a hypothesized ability of phonological awareness, across or even within languages. With multiple measures and different task types or methods, we can formulate rival structural models of trait versus method effects as confirmatory factor models (Eid et al., 2006). Indeed, the wide array of tasks can be distracting, unless researchers were to adopt a latent variable perspective in which each task may potentially be an indicator of an unobserved ability or factor. If the hypothesized ability, such as general phonological awareness, caused performance on the observed tasks, those tasks should relate to each other in a homogeneous manner. Although tasks differ within each language, their similarities may suggest a general ability of phonological awareness in each of these languages. An illustration of the specific implications of such a confirmatory factor model is given in Appendix A.

Such tests of tasks in multiple languages can provide evidence of discriminant validity. Two studies have tested such cross-language models of phonological awareness with confirmatory methods. Branum-Martin et al. (2006) found that at the student

level, after controlling for classroom differences, phonological awareness factors in English and Spanish were statistically separable but highly correlated ($r = .93$). In a study of Cantonese and Mandarin, Chen, Ku, Koyama, Anderson, and Li (2008) found that onset and rime tasks fit a single-factor model across the two languages, but tone awareness did not. With high cross-language correlations, these two studies illustrate the empirical possibility of a confirmatory model of phonological awareness as a single construct (i.e., factor) across languages.

Overall, across the prior reported studies in the meta-analyses, issues of measurement error and structure among these tasks may have been overlooked, leaving a number of questions insufficiently addressed. For example, is high correlation within or across language only due to similarity of task or method? Is low correlation an idiosyncrasy of sample, design, or measure? Does low correlation imply something fundamentally wrong with our notions of how phonological awareness operates in students who speak more than one language?

The current study seeks to answer these questions via confirmatory models fit to the multiple phonological awareness tasks in all of the studies that reported a full correlation matrix for more than two measures of phonological awareness. We investigated the following three research questions, with the details of these models given in the following section:

1. If phonological awareness is a single ability in each language, what does that model suggest about the cross-language correlation of phonological awareness constructs versus cross-language method correlations?
2. To what extent do multiple measures of phonological awareness suggest that it is a single ability across languages, in the presence of method correlations?
3. What do the current models imply with respect to the prior meta-analyses?

Method

Selection and Recording of the Studies

This study analyzes data from the studies included in two recent meta-analyses (Branum-Martin et al., 2012; Melby-Lervåg & Lervåg, 2011). All of the 38 studies reported by Branum-Martin et al. (2012) were examined for whether the study reported more than two measures in a language in a full matrix (i.e., the cross-language correlation for each measure as well as the cross-measure and cross-language correlations). Nineteen of those prior studies reported a full correlation matrix on three or more measures. Three studies had previously been excluded because the measures were not the same across languages (Bursztyn, 1998; Wang, Cheng, & Chen, 2006; Wang, Yang, & Cheng, 2009). The current analysis therefore included 22 studies that reported on 25 samples. These studies are listed in the Appendix B with their sample characteristics and measures.

For each study, the correlation matrix was recorded along with the means and standard deviations. For example, if a study reported five measures in each language, for a total of 10 measures, then the full 10 by 10 correlation matrix was recorded, and such a

study is referred to as a 5×5 . Instead, if the study only had one measure in English and three measures in the other language, the 4 by 4 matrix was recorded. Such a design is called a 1×3 in the current study (see Appendix B). These matrices were used in the latent variable analysis.

Models

We used structural equation models fit to correlations with means and standard deviations (Bollen, 1989; Kline, 2005; MacCallum, Wegener, Uchino, & Fabrigar, 1993), using the sample size reported in each study. With the given correlations, means, and standard deviations, the full covariance matrix can be used to test alternative, confirmatory structural models, as noted in Appendix A (MacCallum et al., 1993). We used Mplus 7 (Muthén & Muthén, 2012) for the models fit to each study in this article. Issues of sample size and fit will be discussed.

Research Question 1 posits that measures indicate one latent factor in each language, with method correlations for measures involving the same task in both languages (see Appendix A). A 4×4 study therefore would have two correlated factors, each with four indicators. Research Question 2 was tested by all the measures indicating only a single factor. Studies that used a different number of outcomes, such as 3×2 , were fit as appropriately reduced versions of these models. Method covariances were included only for observed indicators that were designed to be the same across languages. A factor model of this sort represents the extent to which the given indicators represent the intended construct (factor), with potential residual method covariances. These structures will be presented graphically in the Results section. Research Question 3 was evaluated graphically by summarizing the model results along with estimates from the prior meta-analysis for each language.

Results

The two models were fit to each sample within each study (25 samples from 22 studies), using the sample size reported by the authors (see Appendix B). The two-factor and single-factor tests of the research questions are each reported in turn.

Two-Factor Models: Language-Specific Phonological Awareness (Question 1)

Table 1 shows results for the two-factor models in the 25 samples. The chi-square, comparative fit index (CFI), root-mean-square error of approximation (RMSEA, with 90% confidence interval), and standardized root mean residual (SRMR) are reported for each study. Solely for visual reference, superscripts are used to highlight model indices with a substantial lack of fit ($CFI > .90$, $RMSEA < .06$, and $SRMR < .08$). We do not adhere blindly to these model fit guidelines but discuss sources of misfit and interpretational problems (Marsh, Hau, & Grayson, 2005). Table 1 also contains the estimated latent cross-language correlation between factors of phonological awareness in English and the other language.

Of the 25 models, nine did not obtain interpretable results, indicated by dashes in Table 1. Models with a lack of fit in CFI, RMSEA, or SRMR are marked with a superscript. In the two

Table 1
Fit Statistics for Two-Factor Models (Latent PA in Each Language)

Language and study (sample)	Study notes	χ^2 (df)	<i>p</i>	CFI	RMSEA	90% CI	SRMR	Latent correlation
Greek								
Loizou & Stuart (2003; Greek)	1	—	—	—	—	—	—	—
Loizou & Stuart (2003; British)	2	—	—	—	—	—	—	—
Spanish								
Atwill et al. (2007)	—	2.7 (1)	.10	.99	.16 ^a	[.00, .40]	.02	0.81
Atwill et al. (2010)	—	3.8 (1)	.05	.98	.16 ^a	[.00, .33]	.02	0.76
Branum-Martin et al. (2006)	3	11.9 (5)	.04	1.00	.04	[.01, .07]	.01	0.93
Cisero & Royer (1995, Experiment 1)	—	3.8 (5)	.58	1.00	.00	[.00, .20]	.03	0.89
Cisero & Royer (1995, Experiment 2, Time 1)	1	—	—	—	—	—	—	—
Cisero & Royer (1995, Experiment 2, Time 2)	—	23.8 (5)	<.01	.93	.20 ^a	[.12, .27]	.06	0.77
Leafstedt & Gerber (2005)	—	30.6 (15)	.01	.90	.11 ^a	[.05, .16]	.08	0.98
Gottardo & Mueller (2009)	—	25.3 (12)	.01	.94	.10 ^a	[.04, .15]	.06 ^a	0.66
Bursztyn (1998)	4	—	—	—	—	—	—	—
Korean								
Kim (2009)	—	0.4 (2)	.82	1.00	.00	[.00, .20]	.01	0.89
Wang, Park, & Lee (2006)	—	4.5 (5)	.48	1.00	.00	[.00, .20]	.04	0.94
Cho & McBride-Chang (2005)	5, 6	9.9 (7)	.19	.98	.07 ^a	[.00, .16]	.04	0.93
Cantonese								
Luk (2003)	—	2.4 (6)	.88	1.00	.00	[.00, .11]	.03	0.94
Gottardo et al. (2001)	7	1.8 (3)	.61	1.00	.00	[.00, .17]	.02	0.92
Gottardo et al. (2006)	1, 8	—	—	—	—	—	—	—
McBride-Chang et al. (2006)	1, 5	—	—	—	—	—	—	—
Luk & Bialystok (2008)	1	—	—	—	—	—	—	—
Mandarin								
Wang, Cheng, & Chen (2006)	—	0.3 (2)	.86	1.00	.00	[.00, .13]	.02	0.37
Wang et al. (2005)	4	—	—	—	—	—	—	—
Wang et al. (2009)	—	1.4 (2)	.49	1.00	.00	[.00, .20]	.03	0.06
Yan et al. (2005)	10	13.7 (5)	.02	.88 ^a	.16 ^a	[.06, .27]	.07	0.55
Xu & Dong (2005)	—	18.9 (16)	.27	1.00	.02	[.00, .06]	.02	0.94
Tao et al. (2007)	1	—	—	—	—	—	—	—

Note. Dashes indicate that the model did not yield interpretable results. Study note refers to an explanatory description of the matrix or model used, if needed: (1) The latent correlation estimated >1.0 . (2) The model would not converge, with some modifications suggesting that method correlations were comparable to the English loadings. (3) The student-level matrix from a two-level confirmatory factor model was analyzed. (4) The cross-language and within-language correlations were inconsistent such that no model would converge. (5) The reported correlation matrix was residualized for age in the original article. (6) One-year lag: Korean in kindergarten, English in Grade 1. (7) Sample data consisted of weighted means and standard deviations from four age groups. (8) The reported correlation matrix was residualized for age and level of education in the original article. (9) Covariances were based on weighted means and standard deviations from two age groups. CFI = comparative fit index; RMSEA = root-mean-square error of approximation, with 90% confidence interval (CI); SRMR = standardized root mean residual. The latent correlation is the estimated correlation between the English phonological awareness (PA) factor and the other language PA factor.

^a Indicates a lack of model fit (CFI $< .90$, RMSEA $> .06$, SRMR $> .08$).

Greek samples (Loizou & Stuart, 2003), the two-factor model would not converge, indicating that the reported correlations were inconsistent with the theoretical restrictions implied by the model. The British sample obtained estimates, but a negative residual variance suggested the model was not correct (this study is reexamined in the single-factor results).

Two of the Spanish studies failed to converge, due to heterogeneous correlations inconsistent with the two-factor model (Bursztyn, 1998; Cisero & Royer, 1995, Experiment 2, Time 1). Four of the seven estimable Spanish studies fit well as two-factor models, but the three studies in the lower part of the Spanish section of Table 1 had some evidence of misfit. The estimated latent correlations between English and Spanish were high, ranging from 0.66 to 0.98.

All three two-factor models in Korean fit well, with latent correlations ranging from 0.89 to 0.93 (see Table 1). These high correlations were tested further in Research Question 2.

In Cantonese, only two of the five studies had admissible two-factor models. Both of these models fit well, with latent correlations of 0.92 and 0.94. The three studies with estimation

trouble had latent correlations greater than 1.0, suggesting that a one-factor model could be more appropriate (see the next section).

In Mandarin, four of the six studies had admissible solutions, with three of them fitting well. The latent relations, however, were heterogeneous, ranging from 0.06 to 0.94.

Tests of ■ Single Factor Across Language (Question 2)

In order to test bilingual phonological awareness as a single factor across languages, we also fit a single-factor model to the studies listed in Table 1. Residual method correlations, as in Figure A1 in the Appendix, were still allowed. Table 2 lists the fit statistics for single-factor models for these studies (as in Table 1). In addition, the rightmost column of Table 2 reports the *p* value of the chi-square test of model restriction, comparing the single-factor model to the less restrictive two-factor model in Table 1 (likelihood ratio test, marked with a superscript if $p < .05$). If this chi-square test is above $p = .05$, the fit of the one-factor model is acceptable, compared to the two-factor model.

Table 2

Fit Statistics for Single-Factor Models (PA Is Unitary and Language General)

Language and study (sample)	χ^2 (df)	p	CFI	RMSEA	90% CI	SRMR	p (diff)
Greek							
Loizou & Stuart (2003; Greek)	25.3 (16)	.06	0.88 ^a	0.18 ^a	[.00, .37]	0.07	n/a
Loizou & Stuart (2003; British)	39.4 (16)	<.01	0.66 ^a	0.30 ^a	[.18, .42]	0.15 ^a	n/a
Spanish							
Atwill et al. (2007)	9.5 (2)	.01	0.94	0.24 ^a	[.10, .39]	0.04	0.01 ^a
Atwill et al. (2010)	17.2 (2)	<.01	0.91	0.26 ^a	[.15, .37]	0.05	<.01 ^a
Branum-Martin et al. (2006)	35.7 (6)	<.01	0.99	0.08 ^a	[.06, .10]	0.02	<.01 ^a
Cisero & Royer (1995, Experiment 1)	6.9 (6)	.33	0.99	0.06	[.00, .23]	0.04	0.08
Cisero & Royer (1995, Experiment 2, Time 1)	155.2 (6)	<.01	0.68 ^a	0.50 ^a	[.43, .57]	0.05	n/a
Cisero & Royer (1995, Experiment 2, Time 2)	40.9 (6)	<.01	0.88 ^a	0.24 ^a	[.18, .32]	0.05	<.01 ^a
Leafstedt & Gerber (2005)	30.7 (16)	.01	0.94	0.10 ^a	[.04, .16]	0.08	0.75
Gottardo & Mueller (2009)	47.0 (13)	<.01	0.86 ^a	0.15 ^a	[.11, .20]	0.08	<.01 ^a
Bursztyn (1998)	—	—	—	—	—	—	equivalent
Korean							
Kim (2009)	0.4 (2)	.82	1.00	0.00	[.00, .20]	0.01	equivalent
Wang, Park, & Lee (2006)	5.2 (6)	.52	1.00	0.00	[.00, .18]	0.03	0.40
Cho & McBride-Chang (2005)	10.5 (8)	.23	0.98	0.06	[.00, .14]	0.04	0.44
Cantonese							
Luk (2003)	2.9 (7)	.89	1.00	0.00	[.00, .10]	0.03	0.48
Gottardo et al. (2001)	2.3 (4)	.68	1.00	0.00	[.00, .14]	0.02	0.48
Gottardo et al. (2006)	0.1 (1)	.80	1.00	0.00	[.00, .27]	0.01	n/a
McBride-Chang et al. (2006)	171.6 (1)	<.01	0.62 ^a	0.62 ^a	[.55, .71]	0.16 ^a	n/a
Luk & Bialystok (2008)	37.5 (31)	.20	0.94	0.06	[.00, .12]	0.08	n/a
Mandarin							
Wang, Cheng, & Chen (2006)	0.3 (2)	.86	1.00	0.00	[.00, .13]	0.02	equivalent
Wang et al. (2005)	—	—	—	—	—	—	equivalent
Wang et al. (2009)	1.4 (2)	.49	1.00	0.00	[.00, .20]	0.03	equivalent
Yan et al. (2005)	20.9 (6)	<.01	0.79 ^a	0.20 ^a	[.11, .29]	0.08	0.01 ^a
Xu & Dong (2005)	24.2 (18)	0.11	0.99	0.04	[.00, .07]	0.03	0.07
Tao et al. (2007)	18.4 (11)	0.07	0.93	0.10 ^a	[.00, .17]	0.06	n/a

Note. p(diff) is the p value for the chi-square test of model fit for each study, compared to the respective two-factor model in Table 1. A significant p value (e.g., < .05) in this goodness-of-fit test indicates the restricted single-factor model is significantly worse than the two-factor model. n/a = not applicable, because the two-factor model was not interpretable. equivalent indicates that the single-factor model had the same number of parameters as the two-factor model, and all fit statistics are identical to the two-factor results.

^a Indicates a lack of model fit (comparative fit index < .90, root-mean-square error of approximation > .06, standardized root mean residual > .08, p(diff) < .05).

Only two of the 25 single-factor models were not estimable in Table 2 (Bursztyn, 1998; Wang, Perfetti, & Liu, 2005). Neither of these studies could fit a two-factor model, either (see Table 1). Their covariance structures were heterogeneous, suggesting the theories were not appropriate to these samples.

In Greek, the single-factor model did not fit well for either the Greek or the British subsample. Implications of this lack of fit will be discussed.

In Spanish, three studies had overall reasonable fit for the single-factor model (Branum-Martin et al., 2006; Experiment 1 in Cisero & Royer, 1995; Leafstedt & Gerber, 2005). SRMR did not indicate misfit for any study, but seven of the eight studies had some degree of misfit in RMSEA and three also had a lack of fit in CFI. The test of restriction to a single factor fit for two of the seven models in which a test was possible.

The single-factor model fit well for all three Korean studies. The restriction of two factors down to one factor fit for all three studies (see Table 2), suggesting that the one-factor model was adequate for all three studies of the Korean language.

Four of the five Cantonese studies fit well as single-factor models. The two studies testable against a one-factor model also fit the restriction to a single factor. These single-factor Cantonese

studies are examined for their substantive interpretation in the next section.

In Mandarin, five of the six studies had interpretable results for the single-factor model. Four of the five models with results fit reasonably well. The estimates from these studies are examined next.

Model Results (Question 3)

Overall model fit statistics do not necessarily imply that the model is substantively meaningful. The resulting parameters should be interpreted with respect to their intended roles. In particular, the fully standardized factor loadings are validity coefficients, indexing the correlation between the measure and respective factor: The higher the standardized loading, the better that indicator measures the intended factor (see Appendix A).

The residual correlations indicate the extent to which there is excess shared method variance not explained by the phonological awareness factors in the model. These correlations reflect method effects, or relations due to similarities in the tests and other sources, rather than the factors. The best fitting model for each study is examined next.

Figure 1 presents the single-factor models for the Greek and British subsamples (Loizou & Stuart, 2003). Both models fit poorly (see Table 2). In the British subsample, the factor loadings for English rhyme oddity, onset oddity, and phoneme elision were low (0.26 to 0.48). The loadings for Greek rhyme oddity and phoneme oddity were also low (0.43 and 0.41). Method correlations were high for rhyme oddity (0.73) and phoneme oddity (0.63). Although the fit of the model for the Greek group was poor, the loadings were not disturbingly low. Method correlations for the Greek group shown in Figure 1 were high for phoneme oddity (0.68) and phoneme elision (0.54).

Figure 2 shows the fully standardized results for the best fitting models for Spanish studies. The first three two-factor models in the left-hand column (Atwill, Blanchard, Gorin, & Burstein, 2007; Atwill, Blanchard, Christie, Gorin, & Garcia, 2010; Branum-Martin et al., 2006) all had good validity coefficients. The Branum-Martin et al. (2006) study had low to moderate method correlations (0.13 to 0.41). These three studies all had high latent cross-language correlations. The other two studies that had two-factor results (Experiment 2, Time 2, in Cisero & Royer, 1995;

Gottardo & Mueller, 2009) did not have good model fit (see Table 1). Their validity coefficients were moderate to high, with high cross-language latent correlations (0.77 and 0.66).

The remaining three studies in Figure 2 show the Spanish studies for which the single-factor model fit best. Cisero and Royer's (1995) Experiment 1 fit well, had good loadings, but had some sizable method correlations (0.44 for rhyme detection and 0.32 for final phoneme detection). Cisero and Royer's Experiment 2, Time 1, did not fit well (see Table 2), but had reasonable loadings and one high method correlation (0.64 for rhyme detection). The Leafstedt and Gerber (2005) study fit well but had low validity coefficients for blending and segmenting phonemes (0.30 to 0.47) in each language. The Bursztyn (1998) study did not obtain results and is not shown.

Figure 3 presents the estimated one-factor models for the three studies in Korean. The Korean studies showed moderate to high validity coefficients (0.46 to 0.97) and reasonably low method correlations (0.01 to 0.37).

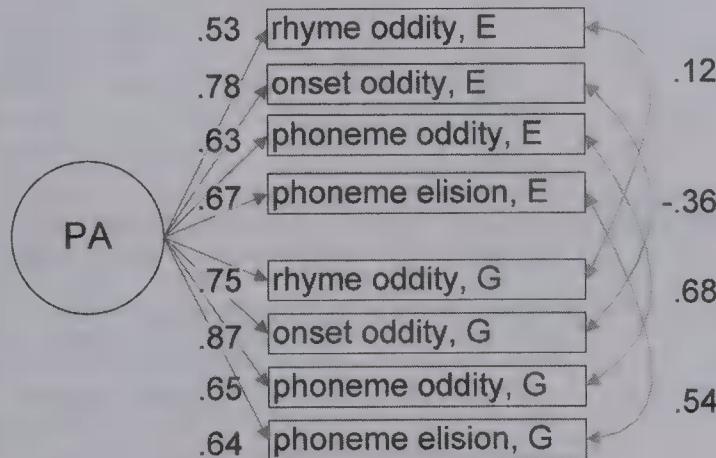
Figure 4 presents the best fitting models for the Chinese studies, with Cantonese on the left side and Mandarin on the right. The single-factor model fit reasonably well for all the Cantonese studies except for one (McBride-Chang, Cheung, Chow, Chow, & Choi, 2006). Despite the reasonable fit for four of the five Cantonese studies, the loadings were only moderate to high (0.10 to 0.92, excluding tone measures) and were particularly low for measures involving tones (0.22 to 0.59). Method correlations were generally moderate. The 5×5 study by Luk and Bialystok (2008) had particularly low loadings, even within languages, suggesting heterogeneous relations among the measures (and therefore a poor match to theory).

The results for Mandarin on the right side of Figure 4 were even more heterogeneous than those for the models for the Cantonese studies. One study did not have dependable estimates (Wang et al., 2005). Two studies did not fit well (Tao, Feng, & Li, 2007; Yan, Yu, & Zhang, 2005) and had only moderate to high loadings (0.36 to 0.88), with low to moderate method correlations (-0.48 to 0.50). In the two 1×3 studies (Wang, Cheng, & Chen, 2006; Wang et al., 2009), the single English measure of phoneme deletion was not a strong indicator (0.37 and 0.06), suggesting that despite model fit, the cross-language nature of the single factor was not well identified. The largest (4×4) Mandarin study (Xu & Dong, 2005) had good fit, good loadings (0.55 to 0.76), and low method correlations (-0.03 to 0.07). Across all Mandarin studies, measures involving tones had fairly uniform validity coefficients in a moderate range (0.39 to 0.48).

Conclusions for Each Language

To graphically summarize the findings and facilitate comparisons, Figure 5 presents the latent variable model estimates of the cross-language correlation along with estimates from the meta-analysis by Branum-Martin et al. (2012). For each language, Figure 5 lists the 25 samples from 22 studies on the left axis. At the bottom of each language section, a square shows the model-based estimate of the cross-language correlation (with 95% CI) from the meta-analysis of all studies that used matched measures (see Table 1 from Branum-Martin et al., 2012). For each study in the present analysis, a circle is shown for the estimated latent cross-language correlation (if estimable) from the two-factor model. The circle is

Loizou & Stuart, 2003, Greek group



Loizou & Stuart, 2003, British group

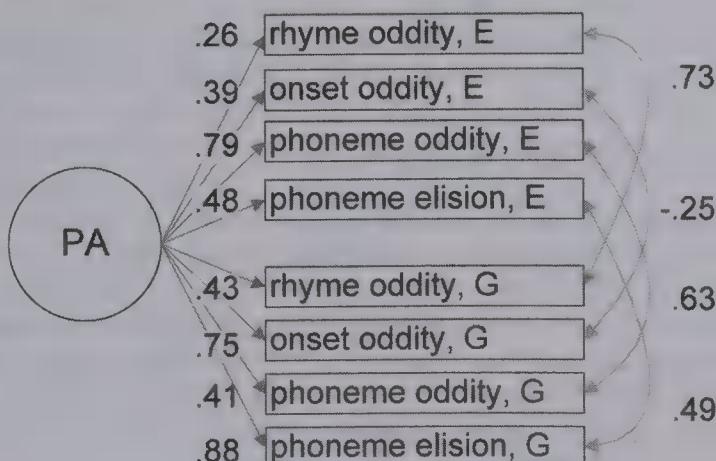


Figure 1. Best fitting results for Greek studies (fully standardized). Neither study in Greek had an estimable two-factor model. Neither study had acceptable fit for one-factor models. Estimates are shown for their validity coefficients and method correlations. PA = phonological awareness; E = English; G = Greek.

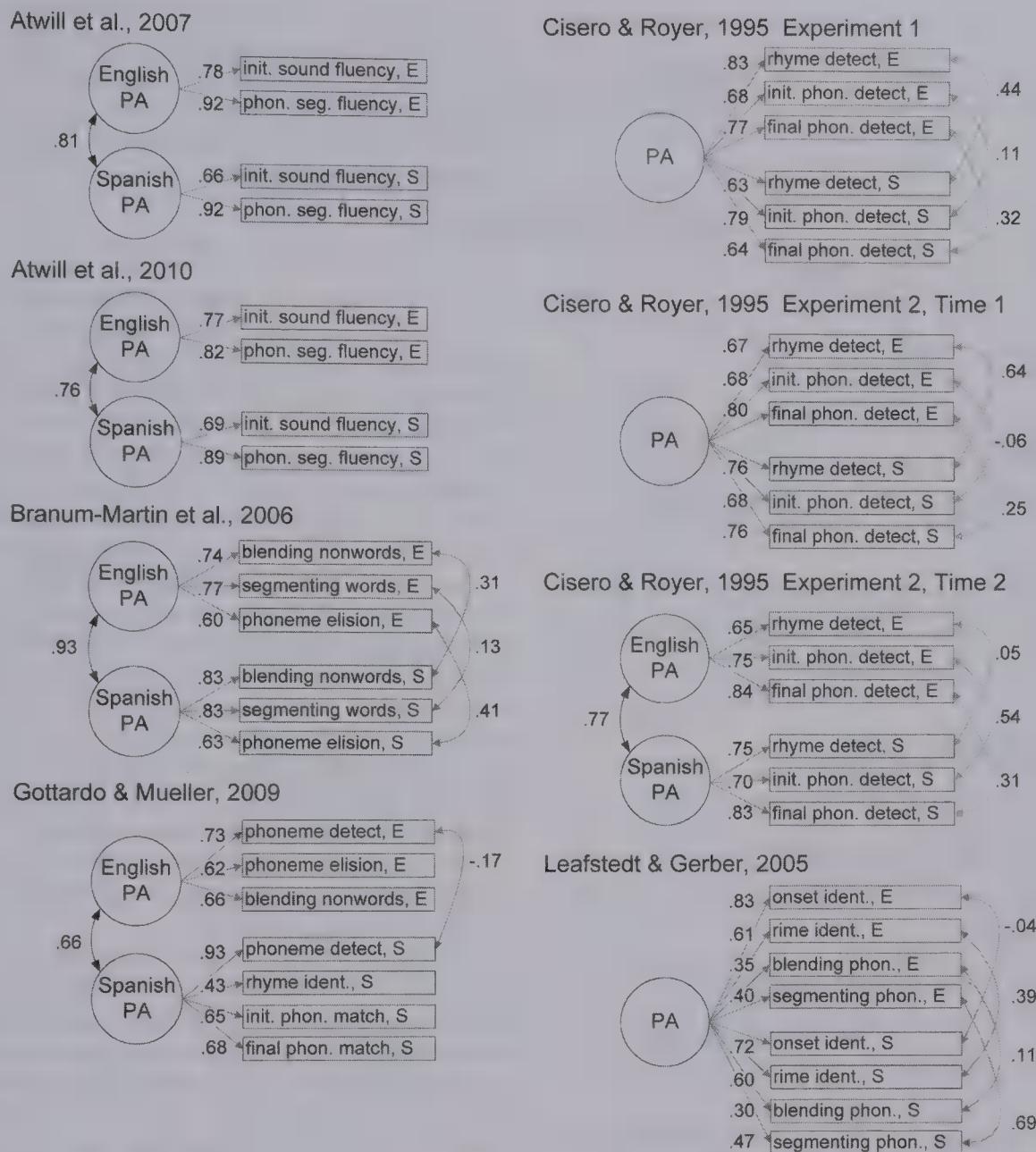


Figure 2. Best fitting results for Spanish studies (fully standardized). Model results are shown for the best fitting two-factor versus one-factor models. PA = phonological awareness; E = English; S = Spanish; init. = initial; phon. = phoneme; seg. = segmenting; ident. = identification; detect. = detection.

filled if the two-factor model had acceptable fit and is empty if the model had two or more indices of substantial misfit (see Table 1). If the single-factor model was estimable, a triangle is shown at 1.0 (filled if the model had acceptable fit and empty if two or more fit indices were outside acceptable range; see Table 2).

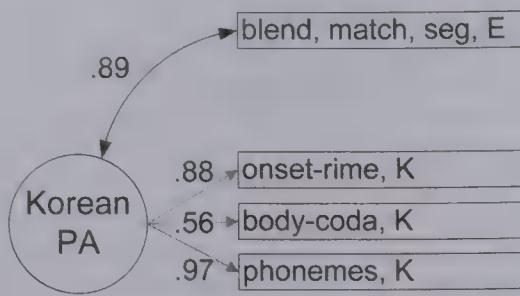
Figure 5 shows that, in Greek, neither sample had a dependable two-factor model (no circles) and the one-factor models did not fit well (empty triangles). The measures were not homogeneously related, with the British group having low validity coefficients and both groups having problematically high method correlations (see Figure 1). It will remain to be seen how consistently Greek phonological awareness tasks correlate in future samples.

In Spanish, the latent correlations shown by the circles in Figure 5 are substantially higher than the estimated meta-analysis correlations for Spanish and English (square). These latent correlations are higher because the measurement error unique to the particular tasks was controlled, yielding disattenuated cross-language relations. These

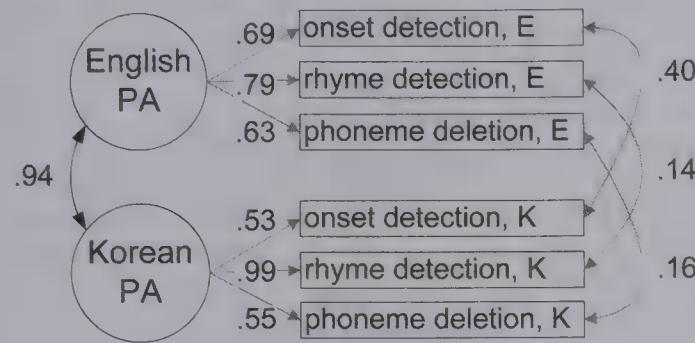
models imply that Spanish and English phonological tasks function relatively well as measures of a consistent construct in each language, with correlations ranging from 0.66 to 0.98. Although the correlations are not quite so high that phonological awareness is a single factor across languages, five of the eight single-factor models had reasonable fit and only two studies passed the restriction to a one-factor model (see Table 2).

The Korean studies shown in Figure 5 show uniformly high correlations and acceptable one-factor solutions. The estimated latent correlations from the two-factor models (black circles in Figure 5) are above the upper bound of the confidence interval for the four Korean studies from the bivariate meta-analysis (square), suggesting the correction for measurement error across multiple tests may be substantial. These high latent correlations and well-fitting single-factor models support the idea that phonological awareness tasks indicate a single factor across the English and Korean languages.

Kim, 2009



Wang, Park et al., 2006



Cho & McBride-Chang, 2005

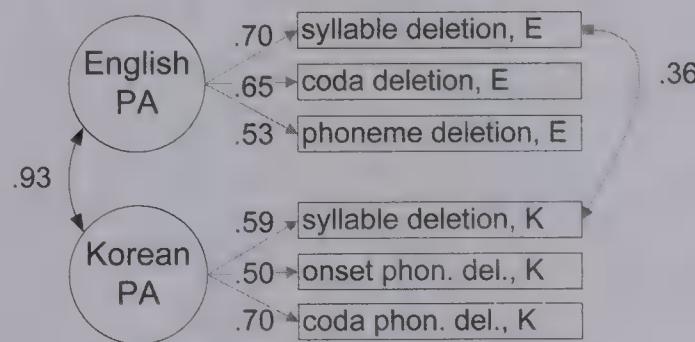


Figure 3. Best fitting results for Korean studies (fully standardized). All three studies of Korean had acceptable fit for their single-factor models. PA = phonological awareness; E = English; K = Korean; blend, match, seg = blending, matching, segmenting; phon. del. = phoneme deletion.

The five Cantonese studies in Figure 5 show that the one-factor model is plausible in terms of fit. However, examination of the low validity coefficients and high method correlations in Figure 4 suggests we should retain some skepticism regarding such a simplistic model of Cantonese and English, especially with respect to measures of tone awareness.

The six Mandarin studies listed in Figure 5 highlight the wide, problematic variability in these studies (also seen in Figure 4). Except for the one large study by Xu and Dong (2005), the latent correlations (circles) were far from the perfect correlations (triangles), even when both models fit. Although two of these studies had only a single measure of English phonological awareness, the low loading for the English measure in these studies (see Figure 4) suggests low convergent validity for English indicating a cross-language latent factor. This lack of convergent validity may indicate limits of the measures or samples or that the model is incorrect: Phonological awareness tasks may function differently in Mandarin than they do in English, yielding two language-specific

factors (perhaps even poorly related). In general, smaller models are less likely to reject and larger models are more difficult to fit, so it will remain to be seen to what extent these mixed results in Mandarin represent issues of specific measures, particular samples, or a problem for our theory of phonological awareness as a coherent ability in Mandarin.

Discussion

The meta-analysis by Branum-Martin et al. (2012) provided a high level of detail on specific design, language, and task effects. However, that analysis focused on only single cross-language correlations nested within each study for measures that were the same in each language. The current study tested the extent to which multiple measures were consistent with a language-specific theory of phonological awareness (two factors) or a language-general theory of phonological awareness (one factor, across languages), all while examining method correlations across similar measures.

The current findings provide important clarification for the previous findings for cross-language correlations of phonological awareness measures (Branum-Martin et al., 2012; Melby-Lervåg & Lervåg, 2011). In particular, measurement error may have played a large role in lowering the correlations estimated in the meta-analyses. Three studies of Korean suggest that the Korean–English correlation among phonological measures adequately represents a single, cross-language factor. In Spanish, phonological tasks are highly related across language but might not represent a single ability.

The findings in Cantonese, suggest the possibility of a high correlation with English for phonological tasks. However, the role of tone awareness is not clear.

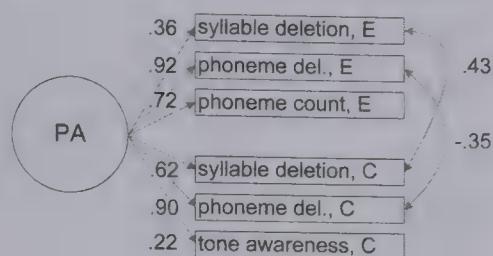
Finally, the nature of the cross-language relation of phonological tasks in Greek and Mandarin is unclear, due to the mixed results across studies. Most of these studies were small, meaning that the correlation matrices used were not necessarily dependable.

In addition, the role of tone awareness in Mandarin is also not clear, but results suggest it is not a strong indicator of phonological awareness (if at all). Tone awareness likely involves other cognitive and linguistic skills not strongly related to phonological awareness as measured by other tasks included in these studies (Chen et al., 2008).

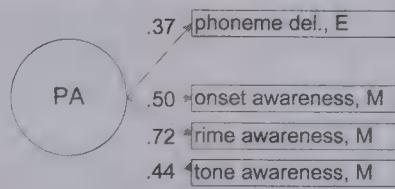
Considerations Concerning Chinese

In Chinese, the factor loadings for syllable tasks were not always similar to the loadings for phoneme-level tasks, which may suggest some task-specific sources of variance related to psycholinguistic grain size (Ziegler & Goswami, 2005, 2006). Children acquiring English need to employ both small and large grain-size strategies to read proficiently (Frost, 1998; Ziegler & Goswami, 2006). Moreover, the Chinese writing system is even less phonologically transparent than the English writing system. It is possible that children learning English along with Mandarin or Cantonese are less able to operate on small grain sizes than are monolingual English children, making phoneme-level tasks not only more difficult but less related to their other phonological skills. Such task-specific differences have been noted in English studies (Anthony et al., 2002; Anthony, Lonigan, Driscoll, Phillips, & Burgess, 2003; Schatschneider, Francis, Foorman, Fletcher, & Mehta, 1999) and Spanish studies (Anthony et al., 2011; Branum-Martin et al., 2006), especially for the more difficult task of segmenting

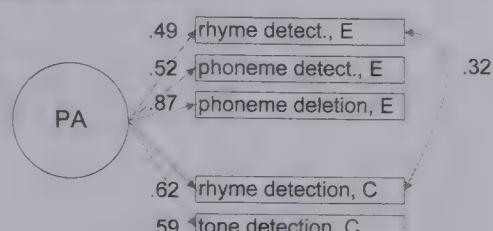
Luk, 2003



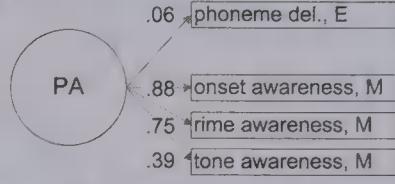
Wang, Cheng, & Chen, 2006



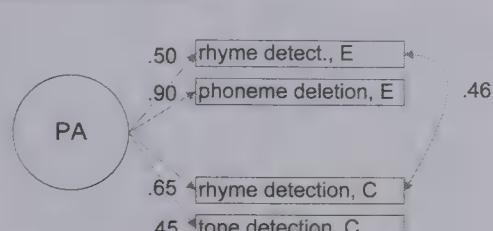
Gottardo et al., 2001



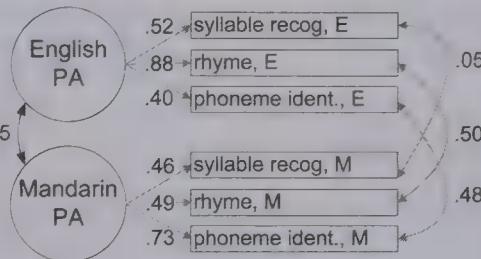
Wang et al., 2009



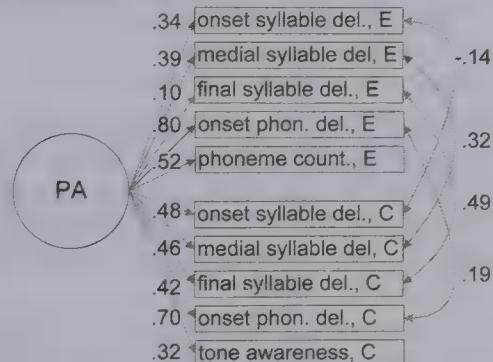
Gottardo et al., 2006



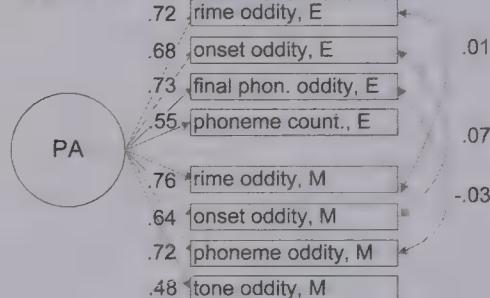
Yan, Yu, & Zhang, 2005



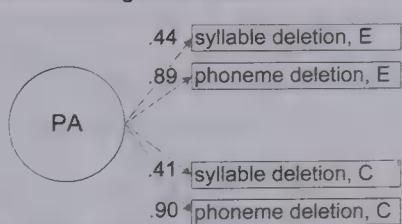
Luk & Bialystok, 2008



Xu & Dong, 2005



McBride-Chang et al., 2006



Tao, Feng, & Li, 2007

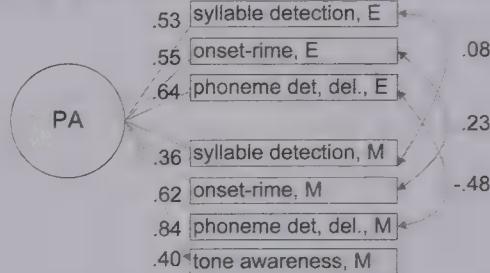


Figure 4. Best fitting results for Chinese studies (fully standardized; Cantonese in left column; Mandarin in right column). Model results are shown for the best fitting two-factor versus one-factor models. One study (Wang et al., 2005) failed to estimate in either model and so is shown with no estimates. PA = phonological awareness; E = English; C = Cantonese; M = Mandarin; del. = deletion; count. = counting; detect. = detection; phon. del. = phoneme deletion; phon. = phoneme; recog. = recognition; ident. = identification.

phonemes. Detailed studies for Chinese speakers are needed at different ages and levels of exposure to English instruction to examine how well phoneme-level tasks operate as indicators of phonological awareness.

In addition, tone awareness might not be as closely aligned with the latent factor as other phonological awareness tasks (Chen et al.,

2008). The extent to which grain size and tonal awareness constitute distinguishable factors or are consistent but weak indicators of a single latent ability will remain to be seen in further discriminant validity studies involving more tasks. Phoneme or tonal tasks in Chinese may have task-specific variability but otherwise could still be valid indicators of a single factor of phonological awareness (in

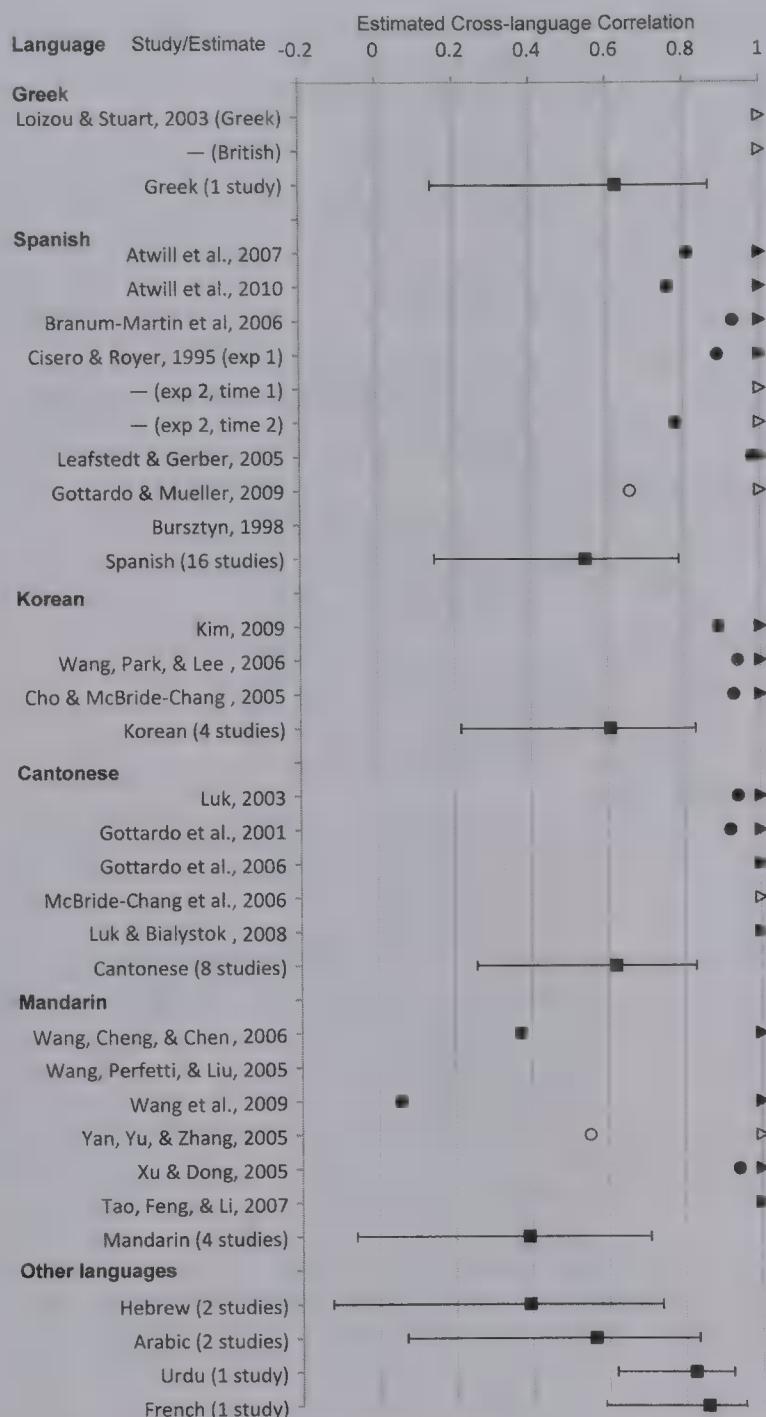


Figure 5. Estimates of cross-language correlation for confirmatory factor models, with estimates from prior meta-analysis. Squares represent the model-based correlation for that language (with 95% confidence interval) from the meta-analysis in Branum-Martin et al. (2012). Each circle represents the estimated latent correlation for the two-factor model for that study. Empty circles indicate the two-factor model had two or more indices of model misfit. Triangles indicate a single-factor model of phonological awareness was estimable for that study—a plausibly perfect correlation—with empty triangles indicating substantial misfit. Studies without a circle or triangle could not be fit with a two- or one-factor model, respectively; exp = experiment.

Chinese or across languages). From the results so far on just 11 studies, there is some support for phonological awareness as a single factor within Chinese, but the role of tone awareness is unclear and potentially different from other tasks involving phonological awareness.

Limitations

Instruction, age, and second versus first language may each be important in activating languages in a bilingual person (Grosjean & Li, 2013), but they were not measured here. Melby-Lervåg and Lervåg (2011) found no effects of instruction upon phonological awareness. Branum-Martin et al. (2012) found cross-language correlations to be slightly lower in older samples. Cross-language effects may differ over time, and they likely influence the attrition and not just the learning of a language (Li, 2013). These complicating factors of development and instruction over time may be responsible for the lack of fit in some samples in the current study. Alternatively, these factors may be responsible for spurious fit of other models in the current sample of studies. Given the difficulty in estimating the effects of instruction and age in prior analyses (Branum-Martin et al., 2012; Melby-Lervåg & Lervåg, 2011), differences due to time, development, and instruction will require a larger base of studies to be evaluated adequately.

Second, this study used English as the anchor language. Other bilingual studies may be similarly revealing, such as between Cantonese and Mandarin (Chen et al., 2008) and between Turkish and Dutch (Verhoeven, 2007). If phonological awareness is indeed universal, we would expect the current results to be clarified in bilingual children in many pairings of languages. Experiments with multilingual children who speak three or more languages can easily be examined in the current framework, either from person-level data or from summary statistics.

Third, because of the complexity of fitting a structural equation model (SEM) to each study, the current study has not explicitly modeled across-study variation (e.g., full meta-SEM; Cheung & Chan, 2005, 2009). More studies might be required, but future examinations could model an across-language latent correlation across studies. Similarly, method correlations in various studies, such as a consistent effect of phoneme-level tasks, could also be directly modeled. Such models could be fit in meta-SEM (Cheung & Chan, 2005, 2009).

Last, the current examination is at the level of summary statistics and only for studies that used bilingual children. Many studies used small samples and were not designed for factor models. Studies within languages at the item level and using detailed linguistic information regarding item features are required to validate these findings (e.g., Anthony et al., 2011). Although interesting and provocative models may be fit to summary statistics, fundamental experiments are needed to validate these models as representative of the underlying cognitive processes. Carefully designed cognitive experiments (e.g., Perfetti et al., 2005; Perfetti & Zhang, 1995) as well as computational linguistic models may be helpful (see reviews by Grosjean & Li, 2013; Thomas & van Heuven, 2005).

Moving Forward in Bilingual Research on Phonological Awareness

The current analyses suggest some changes in thinking about bilingual research, at least for children who may speak more than one language (see Grosjean & Li, 2013; Ziegler & Goswami, 2006). First, researchers should report full descriptive statistics in their studies to allow other researchers to evaluate and reexamine their work (Zientek & Thompson, 2009). Many studies on reading in bilingual samples that used phonological awareness measures

simply did not report cross-language correlations (Branum-Martin et al., 2012). Although several of the studies reported here were primarily concerned with the prediction of reading outcomes, their published correlations allowed us to examine structural questions among the phonological awareness predictors. Second, because researchers design measures to converge or discriminate on certain theoretical constructs in the presence of measurement error, latent variable approaches may help to develop better estimates of these constructs and their relations. Third, only one of the cited studies explicitly modeled classroom level differences. Frequently, students are assigned to classrooms at least partially on the basis of their linguistic abilities, so classroom-level differences are likely to be crucial in bilingual settings (Branum-Martin, Foorman, Francis, & Mehta, 2010; Branum-Martin et al., 2006, 2009). Moreover, classrooms likely differ in their amounts and methods of instruction. We therefore wish to amplify the call for more multivariate, multilevel, and longitudinal examinations of bilingual phenomena (Genesee, Geva, Dressler, & Kamil, 2006).

The current study highlights some crucial questions for future bilingual research in phonological awareness. First, to what extent are various phonological tasks in languages other than English indicative of a single underlying ability? Second, how highly related are phonological abilities across languages, or to what extent can we measure phonological awareness as a general human ability? Third, what are the limiting or facilitating roles of cognitive development, instruction, and the writing system of that language for the nature of phonological awareness and its role in reading?

Conclusion

The current findings give intriguing, if inconsistent empirical support for the idea of phonological awareness being language general. This study extends the findings of the National Literacy Panel on Minority Language Children and Youth (August & Shanahan, 2006) regarding the importance of foundational skills and their cross-language effects. The questions of what kind and how much native language versus target language instruction best facilitates literacy acquisition (and for what kinds of children) will be important to pursue in future research. The current findings provide a basis for approaching such questions by providing a basic framework for conceptualizing the role of phonological awareness across languages. Important extensions will include examining the effects of different instructional models (e.g., transitional, immersion, and dual-language) upon these cross-language effects (Branum-Martin et al., 2012; Melby-Lervåg & Lervåg, 2011). We hope that the current work sparks productive future investigations in this area, leading to improved understanding and delivery of bilingual education.

References

Anthony, J. L., Aghara, R. G., Solari, E. J., Dunkelberger, M. J., Williams, J. M., & Liang, L. (2011). Quantifying phonological representation abilities in Spanish-speaking preschool children. *Applied Psycholinguistics*, 32, 19–49. doi:10.1017/S0142716410000275

Anthony, J. L., Lonigan, C. J., Burgess, S. R., Driscoll, K., Phillips, B. M., & Cantor, B. G. (2002). Structure of preschool phonological sensitivity: Overlapping sensitivity to rhyme, words, syllables, and phonemes. *Journal of Experimental Child Psychology*, 82, 65–92. doi:10.1006/jecp.2002.2677

Anthony, J. L., Lonigan, C. J., Driscoll, K., Phillips, B. M., & Burgess, S. R. (2003). Phonological sensitivity: A quasi-parallel progression of word structure units and cognitive operations. *Reading Research Quarterly*, 38, 470–487. doi:10.1598/RRQ.38.4.3

Atwill, K., Blanchard, J., Christie, J., Gorin, J. S., & Garcia, H. S. (2010). English-language learners: Implications of limited vocabulary for cross-language transfer of phonemic awareness with kindergartners. *Journal of Hispanic Higher Education*, 9, 104–129. doi:10.1177/1538192708330431

Atwill, K., Blanchard, J., Gorin, J. S., & Burstein, K. (2007). Receptive vocabulary and cross-language transfer of phonemic awareness in kindergarten children. *Journal of Educational Research*, 100, 336–346. doi:10.3200/JOER.100.6.336-346

August, D., & Shanahan, T. (Eds.). (2006). *Developing literacy in second-language learners: Report of the National Literacy Panel on language-minority children and youth*. Mahwah, NJ: Erlbaum.

Bollen, K. A. (1989). *Structural equations with latent variables*. New York, NY: Wiley.

Borsboom, D. (2005). *Measuring the mind: Conceptual issues in contemporary psychometrics*. New York, NY: Cambridge University Press.

Branum-Martin, L., Foorman, B. R., Francis, D. J., & Mehta, P. D. (2010). Contextual effects of bilingual programs on beginning reading. *Journal of Educational Psychology*, 102, 341–355. doi:10.1037/a0019053

Branum-Martin, L., Mehta, P. D., Fletcher, J. M., Carlson, C. D., Ortiz, A., Carlo, M. S., & Francis, D. J. (2006). Bilingual phonological awareness: Multilevel construct validation among Spanish-speaking kindergarteners in transitional bilingual education classrooms. *Journal of Educational Psychology*, 98, 170–181. doi:10.1037/0022-0663.98.1.170

Branum-Martin, L., Mehta, P. D., Francis, D. J., Foorman, B. R., Cirino, P. T., Miller, J. F., & Iglesias, A. (2009). Pictures and words: Spanish and English vocabulary in classrooms. *Journal of Educational Psychology*, 101, 897–911. doi:10.1037/a0015817

Branum-Martin, L., Tao, S., Garnaat, S., Bunta, F., & Francis, D. J. (2012). Meta-analysis of bilingual phonological awareness: Language, age, and psycholinguistic grain size. *Journal of Educational Psychology*, 104, 932–944. doi:10.1037/a0027755

Bursztyn, S. B. (1998). *Phonological awareness and reading ability in bilingual native-Spanish and monolingual English-speaking children* (Unpublished doctoral dissertation). Hofstra University.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 56, 81–105. doi:10.1037/h0046016

Chen, X., Ku, Y.-M., Koyama, E., Anderson, R. C., & Li, W. (2008). Development of phonological awareness in bilingual Chinese children. *Journal of Psycholinguistic Research*, 37, 405–418. doi:10.1007/s10936-008-9085-z

Cheung, M. W. L., & Chan, W. (2005). Meta-analytic structural equation modeling: A two-stage approach. *Psychological Methods*, 10, 40–64. doi:10.1037/1082-989X.10.1.40

Cheung, M. W. L., & Chan, W. (2009). A two-stage approach to synthesizing covariance matrices in meta-analytic structural equation modeling. *Structural Equation Modeling*, 16, 28–53. doi:10.1080/10705510802561295

Cho, J.-R., & McBride-Chang, C. (2005). Levels of phonological awareness in Korean and English: A longitudinal study. *Journal of Educational Psychology*, 97, 564–571. doi:10.1037/0022-0663.97.4.564

Cisero, C. A., & Royer, J. M. (1995). The development and cross-language transfer of phonological awareness. *Contemporary Educational Psychology*, 20, 275–303. doi:10.1006/ceps.1995.1018

Durgunoğlu, A. Y., Nagy, W. E., & Hancin-Bhatt, B. J. (1993). Cross-language transfer of phonological awareness. *Journal of Educational Psychology*, 85, 453–465. doi:10.1037/0022-0663.85.3.453

Eid, M., Lischetzke, T., & Nussbeck, F. W. (2006). Structural equation

models for multitrait-multimethod data. In M. Eid & E. Diener (Eds.), *Handbook of multimethod measurement in psychology* (pp. 283–299). Washington, DC: American Psychological Association.

Flege, J. E. (1995). Second language speech learning: Theory, findings, and problems. In W. Strange (Ed.), *Speech perception and linguistic experience: Issues in cross-language research* (pp. 233–277). Timonium, MD: York Press.

Frost, R. (1998). Toward a strong phonological theory of visual word recognition: True issues and false trails. *Psychological Bulletin*, 123, 71–99. doi:10.1037/0033-2909.123.1.71

Genesee, F., Geva, E., Dressler, C., & Kamil, M. (2006). Synthesis: Cross-linguistic relationships. In D. August & T. Shanahan (Eds.), *Developing literacy in second-language learners: Report of the National Literacy Panel on Language-Minority Children and Youth* (pp. 153–174). Mahwah, NJ: Erlbaum.

Gottardo, A., Chiappe, P., Yan, B., Siegel, L. S., & Gu, Y. (2006). Relationships between first and second language phonological processing skills and reading in Chinese-English speakers living in English-speaking contexts. *Educational Psychology*, 26, 367–393. doi:10.1080/01443410500341098

Gottardo, A., & Mueller, J. (2009). Are first- and second-language factors related in predicting second-language reading comprehension? A study of Spanish-speaking children acquiring English as a second language from first to second grade. *Journal of Educational Psychology*, 101, 330–344. doi:10.1037/a0014320

Gottardo, A., Yan, B., Siegel, L. S., & Wade-Woolley, L. (2001). Factors related to English reading performance in children with Chinese as a first language: More evidence of cross-language transfer of phonological processing. *Journal of Educational Psychology*, 93, 530–542. doi:10.1037/0022-0663.93.3.530

Grosjean, F. (2008). *Studying bilinguals*. New York, NY: Oxford University Press.

Grosjean, F., & Li, P. (2013). *The psycholinguistics of bilingualism*. Hoboken, NJ: Wiley.

Gustafsson, J.-E., & Åberg-Bengtsson, L. (2010). Unidimensionality and interpretability of psychological instruments. In S. E. Embretson (Ed.), *Measuring psychological constructs: Advances in model-based approaches* (pp. 97–121). Washington, DC: American Psychological Association.

Kim, Y.-S. (2009). Cross-linguistic influence on phonological awareness for Korean-English bilingual children. *Reading and Writing*, 22, 843–861. doi:10.1007/s11145-008-9132-z

Kline, R. B. (2005). *Principles and practice of structural equation modeling* (2nd ed.). New York, NY: Guilford Press.

Leafstedt, J., & Gerber, M. (2005). Crossover of phonological processing skills: A study of Spanish-speaking students in two instructional settings. *Remedial and Special Education*, 26, 226–235. doi:10.1177/07419325050260040501

Li, P. (2013). Successive language acquisition. In F. Grosjean & P. Li (Eds.), *The psycholinguistics of bilingualism* (pp. 145–167). Hoboken, NJ: Wiley.

Loizou, M., & Stuart, M. (2003). Phonological awareness in monolingual and bilingual English and Greek five-year-olds. *Journal of Research in Reading*, 26, 3–18. doi:10.1111/1467-9817.261002

Luk, G. (2003). *Exploring the latent factors behind inter-language correlations in reading and phonological awareness* (Unpublished master's thesis). York University.

Luk, G., & Bialystok, E. (2008). Common and distinct cognitive bases for reading in English-Cantonese bilinguals. *Applied Psycholinguistics*, 29, 269–289. doi:10.1017/S0142716407080125

MacCallum, R. C., Wegener, D. T., Uchino, B. N., & Fabrigar, L. R. (1993). The problem of equivalent models in applications of covariance structure analysis. *Psychological Bulletin*, 114, 185–199. doi:10.1037/0033-2909.114.1.185

Marsh, H. W., Hau, K.-T., & Grayson, D. (2005). Goodness of fit in structural equation models. In A. Maydeu-Olivares & J. J. McArdle (Eds.), *Contemporary psychometrics: A festschrift for Roderick P. McDonald* (pp. 275–340). Mahwah, NJ: Erlbaum.

Marsh, H. W., Hau, K.-T., & Wen, Z. (2004). In search of golden rules: Comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing Hu and Bentler's (1999) findings. *Structural Equation Modeling*, 11, 320–341. doi:10.1207/s15328007sem1103_2

McBride-Chang, C., Cheung, H., Chow, B. W. Y., Chow, C. S. L., & Choi, L. (2006). Metalinguistic skills and vocabulary knowledge in Chinese (L1) and English (L2). *Reading and Writing*, 19, 695–716. doi:10.1007/s11145-005-5742-x

McDonald, R. P. (1999). *Test theory: A unified treatment*. Mahwah, NJ: Erlbaum.

Melby-Lervåg, M., & Lervåg, A. (2011). Cross-linguistic transfer of oral language, decoding, phonological awareness and reading comprehension: A meta-analysis of the correlational evidence. *Journal of Research in Reading*, 34, 114–135. doi:10.1111/j.1467-9817.2010.01477.x

Muthén, L. K., & Muthén, B. O. (2012). *Mplus user's guide* (7th ed.). Los Angeles, CA: Author.

National Institute of Child Health and Human Development. (2000). *Report of the National Reading Panel: Teaching children to read. An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction: Part I: Phonemic awareness instruction*. Retrieved from <http://www.nichd.nih.gov/publications/nrp/ch2-I.pdf>

Perfetti, C. A. (2003). The universal grammar of reading. *Scientific Studies of Reading*, 7, 3–24. doi:10.1207/S1532799XSSR0701_02

Perfetti, C. A., Liu, Y., & Tan, L. H. (2005). The lexical constituency model: Some implications of research on Chinese for general theories of reading. *Psychological Review*, 112, 43–59. doi:10.1037/0033-295X.112.1.43

Perfetti, C. A., & Zhang, S. (1995). Very early phonological activation in Chinese reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 21, 24–33. doi:10.1037/0278-7393.21.1.24

Perfetti, C. A., Zhang, S., & Berent, I. (1992). Reading in English and Chinese: Evidence for a "universal" phonological principle. In R. Frost & L. Katz (Eds.), *Orthography, phonology, morphology, and meaning* (pp. 227–248). Oxford, England: North-Holland.

Rayner, K., Foorman, B. R., Perfetti, C. A., Pesetsky, D., & Seidenberg, M. S. (2001). How psychological science informs the teaching of reading. *Psychological Science in the Public Interest*, 2, 31–74. doi:10.1111/1529-1006.00004

Schatschneider, C., Francis, D. J., Foorman, B. R., Fletcher, J. M., & Mehta, P. (1999). The dimensionality of phonological awareness: An application of item response theory. *Journal of Educational Psychology*, 91, 439–449. doi:10.1037/0022-0663.91.3.439

Snow, C. E., Burns, M. S., & Griffin, P. (Eds.). (1998). *Preventing reading difficulties in young children*. Washington, DC: National Academy Press.

Tao, S., Feng, Y.-J., & Li, W. (2007). 语音意识的不同成分在汉语儿童英语?读?习中的作用 [The roles of different components of phonological awareness in English reading among Mandarin-speaking children]. *心理发展与教育*, 23, 89–92.

Thomas, M. S. C., & van Heuven, W. J. B. (2005). Computational models of bilingual comprehension. In J. F. Kroll & A. M. B. De Groot (Eds.), *Handbook of bilingualism: Psycholinguistic approaches* (pp. 202–225). New York, NY: Oxford University Press.

Verhoeven, L. (2007). Early bilingualism, language transfer, and phonological awareness. *Applied Psycholinguistics*, 28, 425–439. doi:10.1017/S0142716407070233

Wagner, R. K., & Torgesen, J. K. (1987). The nature of phonological processing and its causal role in the acquisition of reading skills. *Psychological Bulletin*, 101, 192–212. doi:10.1037/0033-2909.101.2.192

Wang, M., Cheng, C., & Chen, S.-W. (2006). Contribution of morphological awareness to Chinese-English biliteracy acquisition. *Journal of Educational Psychology*, 98, 542-553. doi:10.1037/0022-0663.98.3.542

Wang, M., Park, Y., & Lee, K. R. (2006). Korean-English biliteracy acquisition: Cross-language phonological and orthographic transfer. *Journal of Educational Psychology*, 98, 148-158. doi:10.1037/0022-0663.98.1.148

Wang, M., Perfetti, C. A., & Liu, Y. (2005). Chinese-English biliteracy acquisition: Cross-language and writing system transfer. *Cognition*, 97, 67-88. doi:10.1016/j.cognition.2004.10.001

Wang, M., Yang, C., & Cheng, C. (2009). The contributions of phonology, orthography, and morphology in Chinese-English biliteracy acquisition. *Applied Psycholinguistics*, 30, 291-314. doi:10.1017/S0142716409090122

Xu, F., & Dong, Q. (2005). 汉语儿童汉语与英语语音意识发展的关系 [The relationship between development of Chinese and English phonological awareness in primary school]. *心理发展与教育*, 21, 31-35.

Yan, R., Yu, G., & Zhang, L. (2005). 双语儿童语音意识与词认?关系的研究 (The relationship between phonological awareness and English word reading among Chinese kindergarten children). *心理科学*, 28, 304-307.

Ziegler, J. C., & Goswami, U. (2005). Reading acquisition, developmental dyslexia, and skilled reading across languages: A psycholinguistic grain size theory. *Psychological Bulletin*, 131, 3-29. doi:10.1037/0033-2909.131.1.3

Ziegler, J. C., & Goswami, U. (2006). Becoming literate in different languages: Similar problems, different solutions. *Developmental Science*, 9, 429-436. doi:10.1111/j.1467-7687.2006.00509.x

Zientek, L. R., & Thompson, B. (2009). Matrix summaries improve research reports: Secondary analyses using published literature. *Educational Researcher*, 38, 343-352. doi:10.3102/0013189X09339056

Appendix A

Example of a Confirmatory Factor Model

To illustrate how multiple tasks can be used to test hypotheses of measuring trait versus method effects, we present a conceptual example with a specified model. Consider an example study in which three tasks are given to children in English (e.g., blending, segmenting, and elision) and three tasks are given in another language (e.g., blending, segmenting, and sound matching).

Figure A1 presents a structural equation model diagram of a confirmatory factor model for these three tasks in two languages with fully standardized results. Circles represent latent factors: the number and type of constructs (traits) we wish to test. Rectangles represent observed test scores. Straight arrows represent measurement relations (pattern coefficients or loadings). Curved, double-headed arrows represent correlations. Gray curved, double-headed arrows represent variance (standardized to 1.0).

Overall, this model implies a particular pattern of correlations among variables: Correlations are caused by the proposed factors, with method correlations, each of which can be evaluated. Where

measures are the same in both languages, a residual method correlation can be included. If measures are not the same across languages, the factor model implies that correlations among tests are caused only by the hypothesized factors. This sort of model has four important characteristics for questions of cross-language phonological awareness:

1. Overall fit of the model represents the extent to which the data fit this theoretically specified model. That is, indices of model fit suggest how closely the covariance structure implied by the theoretical model matches the actual covariance reported for the study. Good model fit supports validity of the theoretically specified constructs, and rejection of the model suggests that our theory for the number and pattern of constructs is not correct (Borsboom, 2005; Marsh et al., 2005; Marsh, Hau, & Wen, 2004).
2. Factor loadings (λ) represent the sensitivity or quality of that test for measuring the proposed latent factor (Bollen, 1989; Gustafsson & Åberg-Bengtsson, 2010; McDonald, 1999). The squared loadings from the fully standardized solution represent the proportion of variance in the test explained by the latent factor (R^2).
3. The cross-language correlation between factors (curved, double-headed arrow on the left side) represents the relation between latent constructs, corrected for measurement error. If this correlation is high, it could suggest a more parsimonious model of only one latent factor across languages. A model of a single, cross-language factor can also be tested.

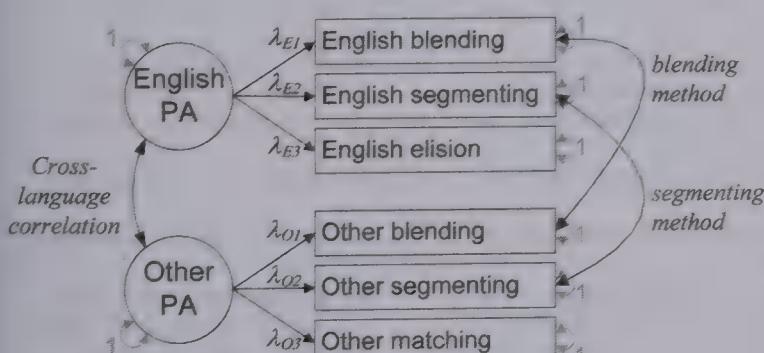


Figure A1. Specification of a cross-language model of phonological awareness (fully standardized). Mean structure not shown. PA = phonological awareness.

4. The method correlations (right side) represent residual relations due to that specific method (e.g., blending) that is not predicted by the common factor of phonological awareness in that language. The extent

to which these method correlations are higher than the loadings or the latent correlation suggests that method effects may be more important than trait effects.

Appendix B

Table of Study Characteristics

Study and language	Measures	Sample	English	Other
Greek				
Loizou & Stuart (2003)	6 × 6	<i>n</i> = 18 in Greece; 16 in UK; pre-kindergarten	rhyme oddity, syllable completion, onset oddity, initial phoneme identification, single phoneme onset oddity, phoneme elision	rhyme oddity, syllable completion, onset oddity, initial phoneme identification, single phoneme onset oddity, phoneme elision
Spanish				
Atwill et al. (2007)	2 × 2	<i>n</i> = 68; kindergarten, southwestern US	initial sound fluency, phoneme segmentation fluency	initial sound fluency, phoneme segmentation fluency
Atwill et al. (2010)	2 × 2	<i>n</i> = 68; kindergarten, southwestern US	initial sound fluency, phoneme segmentation fluency	initial sound fluency, phoneme segmentation fluency
Branum-Martin et al. (2006)	3 × 3	<i>n</i> = 812; kindergarten, southwestern US	blending nonwords, segmenting words, phoneme elision	blending nonwords, segmenting words, phoneme elision
Cisero & Royer (1995)	3 × 3	<i>n</i> = 36–99; kindergarten–first grade, northeastern US	rhyme detection, initial phoneme detection, final phoneme detection	rhyme detection, initial phoneme detection, final phoneme detection
Leafstedt & Gerber (2005)	4 × 4	<i>n</i> = 89; first grade, western US	onset, rime, blending, segmenting	onset, rime, blending, segmenting
Gottardo & Mueller (2009)	3 × 4	<i>n</i> = 114; Grades 1–2, Canada	phoneme detection, phoneme elision, blending nonwords	phoneme detection, rhyme identification, initial phoneme matching, final phoneme matching
Bursztyn (1998)*	1 × 3	<i>n</i> = 45; Grades 1–2, northeastern US	quick rhyming	segmenting, blending, initial phoneme matching
Korean				
Kim (2009)	1 × 3	<i>n</i> = 33; kindergarten, eastern and western US	total of blending, matching, segmenting	blending and segmenting: rimes, body-coda, phonemes
Wang, Park, & Lee (2006)	3 × 3	<i>n</i> = 45; Grades 1–3, eastern US	onset detection, rhyme detection, phoneme deletion	onset detection, rhyme detection, phoneme deletion
Cho & McBride-Chang (2005)	3 × 3	<i>n</i> = 91; Grade 2, Korea	syllable deletion, coda deletion, phoneme deletion	syllable deletion, onset phoneme deletion, coda phoneme deletion
Cantonese				
Luk (2003)	3 × 3	<i>n</i> = 33; Grade 1, Canada	syllable deletion, phoneme onset deletion, phoneme counting	syllable deletion, phoneme onset deletion, tonal awareness
Gottardo et al. (2001)	3 × 2	<i>n</i> = 65; Grades 1–8, Canada	rhyme detection, phoneme detection, phoneme deletion	rhyme detection, tone detection
Gottardo et al. (2006)	2 × 2	<i>n</i> = 40; Grades 1–8, Canada	rhyme detection, phoneme deletion	rhyme detection, tone detection
McBride-Chang et al. (2006)	2 × 2	<i>n</i> = 217; kindergarten, Hong Kong	syllable deletion, phoneme onset deletion	syllable deletion, phoneme onset deletion
Luk & Bialystok (2008)	5 × 5	<i>n</i> = 57; Grade 1, Canada	onset syllable deletion, medial syllable deletion, final syllable deletion, onset phoneme deletion, phoneme counting	onset syllable deletion, medial syllable deletion, final syllable deletion, onset phoneme deletion, tone awareness
Mandarin				
Wang, Cheng, & Chen (2006)*	1 × 3	<i>n</i> = 64; Grades 1–5, eastern US	phoneme deletion	onset awareness, rime awareness, tone awareness
Wang et al. (2005)	3 × 3	<i>n</i> = 46; Grades 2–3, eastern US	onset matching, rime matching, phoneme deletion	onset matching, rime matching, tone matching
Wang et al. (2009)*	1 × 3	<i>n</i> = 78; Grade 1, eastern US	phoneme deletion	onset awareness, rime awareness, tone awareness
Yan et al. (2005)	3 × 3	<i>n</i> = 64; kindergarten, China (Beijing)	syllable recognition, rhyme, phoneme identification	syllable recognition, rhyme, phoneme identification

Appendix B (continued)

Study and language	Measures	Sample	English	Other
Xu & Dong (2005)	<i>4 × 4</i>	<i>n</i> = 302; Grades 1, 3, 5, Beijing	rime oddity, onset oddity, final phoneme oddity, phoneme counting	rime oddity, onset oddity, phoneme oddity, tone oddity
Tao et al. (2007)	<i>3 × 4</i>	<i>n</i> = 74; Grades 3 and 5, Beijing	syllable detection and deletion, onset-rime detection and deletion, phoneme detection and deletion	syllable detection and deletion, onset-rime detection and deletion, phoneme detection and deletion, tone detection and substitution

Note. An asterisk indicates a study excluded from the previous meta-analysis (Branum-Martin et al., 2012) for not having matched measures across language. Italics indicate a measure that is matched across languages. The Measures column shows the number of measures in English first and the other language second, so that *1 × 3* indicates that correlations were given for 1 measure in English and 3 measures in the other language. UK = United Kingdom; US = United States.

Received December 4, 2012

Revision received April 28, 2014

Accepted May 3, 2014 ■

Literacy Skill Development of Children With Familial Risk for Dyslexia Through Grades 2, 3, and 8

Kenneth Eklund, Minna Torppa, Mikko Aro, Paavo H. T. Leppänen, and Heikki Lyytinen
University of Jyväskylä

This study followed the development of reading speed, reading accuracy, and spelling in transparent Finnish orthography in children through Grades 2, 3, and 8. We compared 2 groups of children with familial risk for dyslexia—1 group with dyslexia (Dys_FR, $n = 35$) and 1 group without (NoDys_FR, $n = 66$) in Grade 2—with a group of children without familial risk for dyslexia (controls, $n = 72$). The Dys_FR group showed persistent deficiency, especially in reading speed, and, to a minor extent, in reading and spelling accuracy. The Dys_FR children, contrary to the other 2 groups, relied heavily on letter-by-letter decoding in Grades 2 and 3. In children not fulfilling the criteria for dyslexia in Grade 2, the familial risk did not substantially affect the subsequent development of literacy skills.

Keywords: reading development, spelling development, familial risk, dyslexia, longitudinal

Literacy skills are a key to educational and occupational success in most societies. For a considerable proportion of the population, difficulties in reading and spelling development make them vulnerable to underachievement throughout their school years and even beyond (Snowling, Adams, Bishop, & Stothard, 2001). Children with a family history of dyslexia represent a substantial part of this population: 34%–66% of children born to families with dyslexia have been reported to have severe difficulties in reading and spelling acquisition during the first grades at school (Pennington & Lefly, 2001; Puolakanaho et al., 2007; Scarborough, 1990; Snowling, Callagher, & Frith, 2003). The majority of studies of reading development have focused on reading accuracy, and less is known about the development of reading speed (Landerl & Wimmer, 2008; Share, 2008) and spelling (Lervåg & Hulme, 2010). In studies of reading speed, a few longitudinal follow-ups have spanned beyond Grade 3 (de Jong & van der Leij, 2003; Landerl & Wimmer, 2008; Parrila, Aunola, Leskinen, Nurmi, & Kirby, 2005), but follow-ups at school age with samples including children with familial risk for dyslexia are scarce (see, however, Snowling, Muter, & Carroll, 2007; van Bergen et al., 2011). This longitudinal study examines reading and spelling development across Grades 2, 3, and 8 in three groups: children with familial risk for dyslexia and dyslexia in Grade 2, children with familial risk but no dyslexia in Grade 2, and children without a familial risk and without dyslexia. We had three aims: (a) to study the stability of reading and spelling skills beyond the literacy acquisition phase,

(b) to examine the effect of familial risk on reading and spelling development, and (c) to examine the effect of reading task and material (word list, text, and pseudoword text) on reading speed in different groups at different ages.

Stability in Reading Speed, Reading Accuracy, and Spelling

Only a few studies have described reading and spelling development from childhood to adolescence in a longitudinal design, and most of them have involved English-speaking children and have focused on the development of reading accuracy (Francis, Shaywitz, Stuebing, Shaywitz, & Fletcher, 1996; Parrila et al., 2005; Shaywitz et al., 1995). During recent years, reading speed and fluency (speed adjusted for accuracy) have begun to receive more attention in developmental reading research. In one of the few studies focusing on the development of reading speed, Landerl and Wimmer (2008) reported high stability and steady growth in a sample of German-speaking (Austrian) children in Grades 1, 4, and 8. Correlations between reading speed measures at different grade levels varied from .59 to .81, indicating high stability, which was confirmed at the individual level: eight out of 11 slow readers in Grade 1 were still at least one standard deviation below the sample average in Grade 8. Similarly, high correlations were reported in reading speed between words (.69) and nonwords (.66) in a shorter Dutch follow-up ranging from Grades 1 to 3 (de Jong & van der Leij, 2002) as well as in English between Grades 1 and 2 in word list (.79) and oral text reading (.82) fluency (Kim, Wagner, & Lopez, 2012). In Finnish, correlations between Grade 1 (fall) and Grade 2 (spring) have varied from .59 in text reading fluency (Parrila et al., 2005) to .67 in word recognition fluency (Torppa et al., 2007).

The stability of reading accuracy has also been reported to be high. In an English-speaking Canadian sample, the across-grade correlations varied between .47 and .94 in the yearly assessments from Grades 1 to 5 (Parrila et al., 2005). In transparent orthographies, the development of reading accuracy is very different from English, because the acquisition of reading accuracy in transparent

This article was published Online First June 16, 2014.

Kenneth Eklund, Department of Psychology, University of Jyväskylä; Minna Torppa, Department of Teacher Education, University of Jyväskylä; Mikko Aro, Department of Education, University of Jyväskylä; Paavo H. T. Leppänen and Heikki Lyytinen, Department of Psychology, University of Jyväskylä.

Correspondence concerning this article should be addressed to Kenneth Eklund, Department of Psychology, University of Jyväskylä, P.O. Box 35, Agora, 40014 Jyväskylä, Finland. E-mail: Kenneth.M.Eklund@jyu.fi

orthographies is fast. In a cross-language comparison of seven languages, Aro and Wimmer (2003) reported that the percentage of accurately read pseudowords approached 90% at the end of Grade 1 in all six orthographies (German, Dutch, Swedish, French, Spanish, and Finnish) other than English. Even children with dyslexia have been reported to read at least words with high accuracy after Grade 1: the average accuracy percentage was 91% in a Dutch sample of children with dyslexia (de Jong & van der Leij, 2003). Therefore, reading accuracy is seldom followed up and reported on in transparent orthographies after Grade 1. U. Leppänen, Niemi, Aunola, and Nurmi (2006) have, however, reported moderate to high correlations, ranging from .52 to .91, in reading accuracy of words and sentences in a Finnish sample in four assessments during Grades 1 and 2.

As noted in various definitions of dyslexia, including the one from the International Dyslexia Association, problems in spelling are one key marker of dyslexia: "Dyslexia is a specific learning disability . . . characterized by difficulties with accurate and/or fluent word recognition and by poor spelling and decoding abilities" (p. 2; Lyon, Shaywitz, & Shaywitz, 2003). Spelling development, however, has attracted less attention (Caravolas, Hulme, & Snowling, 2001; Lervåg & Hulme, 2010). There are studies that have examined the early prerequisites and predictors of spelling skill during the early grades of school in different orthographies (e.g., Furnes & Samuelsson, 2010; Kim & Petscher, 2011; U. Leppänen et al., 2006; Torppa et al., 2013; Wimmer & Mayringer, 2002). But there are only a few longitudinal follow-ups that have examined the stability of spelling skill beyond the first grades at school in children without dyslexia (Abbot, Berninger, & Fayol, 2010; Landerl & Wimmer, 2008; Lervåg & Hulme, 2010), and with dyslexia (Shaywitz et al., 1999; Snowling et al., 2007). Several studies have shown that children with reading difficulties are often poor in both reading and spelling (de Jong & van der Leij, 2003; Pennala et al., 2010; Pennington & Lefly, 2001; Puolakanaho et al., 2008; van Bergen et al., 2012). In addition, the finding that spelling training in children with dyslexia enhances reading skills supports the idea of a close relationship between reading and spelling (Ise & Schulte-Körne, 2010). However, dissociation between spelling and reading has also been reported (Fayol, Zorman, & Lété, 2009; Moll & Landerl, 2009; Wimmer & Mayringer, 2002). Reported correlations between two assessments of spelling have indicated moderate to high stability in English (.62–.92, in Grades 1–7; Abbott et al., 2010), in Norwegian (.47–.78, Grades 1–3; Lervåg & Hulme, 2010), and in German (.44–.77; Landerl & Wimmer, 2008). A tendency for stronger correlations between words compared with pseudowords (.67–.78 vs. .47–.59, respectively; Lervåg & Hulme, 2010) as well as later versus earlier grades (.44–.47 in Grades 1–4 vs. .77 for Grades 4 and 8; Landerl & Wimmer, 2008) has also been reported. Correlational stability was not reported in the studies with a sample of children with dyslexia, but stable group differences between children with and without dyslexia were found between Grades 6 and 9 (ages 9–14 years; Shaywitz et al., 1995) and between 8 and 12 years of age (Snowling et al., 2007). Our study is, to our knowledge, the first to examine the stability in reading speed and accuracy, as well as in spelling, in a sample of children with and without familial risk for dyslexia across a long time period from Grade 2 to Grade 8 (ages 8–14 years).

Familial Risk as a Continuum

Several candidate susceptibility genes have been found to be linked to developmental dyslexia (Galaburda, LoTurco, Ramus, Fitch, & Rosen, 2006; Giraud & Ramus, 2012; Scerri & Schulte-Körne, 2010), and the idea of multiple risk factors, some of which are transmitted also to offspring without dyslexia, is widely accepted (Bishop, 2009; Pennington, 2006; Pennington & Lefly, 2001; Pennington et al., 2012; Snowling, 2008; Snowling et al., 2003). Pennington (2006) has suggested that multiple risk factors both in the genome and environment lead to a continuum of vulnerability instead of a dichotomous distribution of risk. At the behavioral level, this suggestion has been tested by comparing the performance of children with familial risk, either with or without dyslexia, and controls. If the familial risk is continuous, the group of children with familial risk but no dyslexia also should show lower performance in the underlying cognitive skills (endophenotypes) compared with controls. The studies comparing these three groups have mainly shown that children with familial risk who do not fulfill the criteria of dyslexia perform significantly below the level of the controls in certain language and literacy skills both prior to and after school entry (Boets et al., 2010; Gallagher, Frith, & Snowling, 2000; Pennington & Lefly, 2001; Snowling, 2008; Snowling et al., 2003; van Bergen et al., 2011, 2012).

In English-speaking children, Pennington and Lefly (2001) found that the scores of children with familial risk but without dyslexia in Grade 2 were significantly lower—on average, 0.5 of a standard deviation—than the scores of children with no familial risk and no dyslexia in all except one reading task. In line with this result, Snowling et al. (2003) found that the at-risk children without dyslexia showed poor performance in nonword reading and phonetic spelling at the age of 6 years and poor skills in spelling, nonword reading accuracy, and reading comprehension at the age of 8 years. In addition, in a follow-up study in adolescence Snowling et al. (2007) reported that the at-risk unimpaired children had weaker performance than controls in exception word reading, text reading accuracy, and all timed reading tasks. However, the at-risk unimpaired children did not show deficient performance in word reading accuracy at 8 years (Snowling et al., 2003), either in untimed nonword reading accuracy or in reading comprehension in adolescence (Snowling et al., 2007). The classification of children with dyslexia in these studies was based on a composite score including word reading and spelling accuracy as well as reading comprehension (Snowling et al., 2003, 2007).

Van Bergen et al. (2011, 2012) have also found evidence for the continuity of genetic liability of dyslexia in Dutch samples of children. At the end of Grade 2, children with familial risk but no dyslexia scored higher than children with familial risk and dyslexia but were impaired, compared with controls, in all literacy measures (i.e., reading accuracy and fluency), and spelling (van Bergen et al., 2012). It is noteworthy that the differences between the groups were similar irrespective of whether the items were words or nonwords. In another Dutch sample (van Bergen et al., 2011), where dyslexia was diagnosed in Grade 5, the at-risk nondyslexic children performed worse in nonword reading fluency in Grades 1, 2, and 5 than typically reading control children. However, in word reading fluency, the groups did not differ anymore in Grade 5. The classification of children with dyslexia was based solely on reading fluency (van Bergen et al., 2011, 2012).

On the other hand, the Dutch-speaking sample in Boets et al. (2010) showed support for the continuous nature of the effects of familial risk in prereading skills before school age only, but not in literacy skills at school age. Boets et al. (2010) found that the non-dyslexic at-risk children were poorer than control children in nonword repetition at kindergarten but not in Grades 1 and 3. They also found that this group was as good as the control group in word reading accuracy and speed as well as in nonword reading speed in Grades 1 and 3. The only significant differences found between these two groups at school age were in nonword reading accuracy and spelling, both of which emphasize accurate decoding ability (Boets et al., 2010). In a Finnish sample, no significant differences were found between children with familial risk but without dyslexia and typical readers from control families in reading-related prereading skills, including language skills and phonological sensitivity at age 1 year 6 months through age 5 years 6 months and rapid serial naming and letter knowledge at age 3 years 6 months through age 5 years 6 months (Torppa, Lyytinen, Erskine, Eklund, & Lyytinen, 2010). In Grade 2, the same groups did not differ from each other in reading accuracy or speed or in spelling, irrespective of whether the material was individually presented words or nonwords, or presented in the form of a list or text (Torppa et al., 2010). However, differences among the same three groups were found in brain responses to nonspeech pitch change in sounds at birth (P. H. T. Leppänen et al., 2010) as well as in the ability to discriminate speech stimuli with a barely perceivable difference in Grade 2 and in Grade 3 (Pennala et al., 2010). In the sample, the classification of dyslexia was based on reading speed and accuracy as well as on spelling accuracy (Pennala et al., 2010; Torppa et al., 2010).

Because several factors vary between these studies (e.g., language and orthography, age and way of classifying dyslexia, stimuli and tasks used), it is difficult to draw firm conclusions of the reasons for differing findings. It seems, however, that differences among the groups without reading difficulty and with or without familial risk are more clearly present early in the development of skills (Boets et al., 2010; Snowling et al., 2007; van Bergen et al., 2010). In addition, typical readers with or without familial risk have performed at the same level in tasks such as word reading and reading comprehension, where it is possible to make use of abilities other than phonological-decoding-related skills (i.e., semantic and syntactic skills and contextual cues) to facilitate reading (Boets et al., 2010; Snowling, 2007; Snowling et al., 2008; van Bergen et al., 2011) or when there is less pressure, such as no time limit (Snowling et al., 2007). Based on the previous findings from Grade 2 in the Finnish sample (Torppa et al., 2010) and the fact that group differences tend to diminish along with age (Boets et al., 2010; Snowling et al., 2007; Torppa et al., 2010; van Bergen et al., 2011), we expected that children with familial risk but no reading difficulty in Grade 2 would not differ from the control children in any of the reading and spelling measures in Grades 3 and 8.

The Effect of Task on Reading Speed

Differences in reading speed across tasks have been interpreted to reflect different processes involved in different reading tasks. Reading pseudowords has generally been considered as a good measure of decoding ability because it requires grapheme-to-phoneme decoding (e.g., Coltheart, Rastle, Perry, Langdon, &

Ziegler, 2001). Children with reading difficulty have been shown to have serious deficiency with this type of decoding, at least in opaque orthographies (Bergmann & Wimmer, 2008; Ziegler, Perry, Ma-Wyatt, Ladner, & Shulte-Körne, 2003). In word reading, whether presented in a list or text format, the use of lexicon (i.e., activation of lexical representations) can substantially quicken reading speed by enabling fast whole-word recognition (Coltheart et al., 2001; Frith, Wimmer, & Landerl, 1998).

Reading time of dyslexic readers has been shown to be more dependent on word length both in pseudoword and word reading than in control children (Ziegler et al., 2003; Zoccolotti et al., 2005). These findings have been interpreted to support the view that dyslexic readers rely more on phonological letter-by-letter decoding than typical readers. On the other hand, Bergmann and Wimmer (2008) have shown that even dyslexic readers (German speaking, ages 15–18 years) rely on the direct access to lexical information when reading from print to phonology for familiar letter strings, even though they are slower than nonimpaired readers. The so-called *lexicality effect* (i.e., the faster reading of word stimuli compared with reading nonwords), has been demonstrated to increase with grade level from Grade 1 to Grade 5 (Zoccolotti, De Luca, Di Filippo, Judica, & Martelli, 2009). This finding has been interpreted to be a result of more efficient use of the lexical information as children get older (Zoccolotti et al., 2009). This gradual shift from mainly using sequential letter-to-sound decoding to the predominant use of fast whole-word recognition during the development of reading acquisition gets support from Vaessen and Blomert (2010). Their study shows increasing speed differences over years (Grades 1–6) between word and pseudoword reading. In the present study, we examined whether in Finnish, (similar to the case in Italian; Zoccolotti et al., 2009), children with dyslexia show a later developmental shift of emphasis from phonological decoding strategy to lexical processing than typically reading children. We assessed this shift by comparing speed in pseudoword text reading to word list and text reading in Grades 2, 3, and 8.

Skilled fluent reading is based on accurate and automatic word recognition in different contexts that facilitates the activation of semantic processes. Together with the appropriate use of prosody, reading fluency supports quick comprehension of reading material (Kuhn, Schwanenfugel, & Meisinger, 2010). Words in context are usually read faster and more accurately than the same words without context (Jenkins, Fuchs, van den Broek, Espin, & Deno, 2003). According to Posner and Snyder (1975), there are two processes used for speeding up word identification in a textual context: automatic semantic activation of lexical memory and slow-acting attention-demanding conscious use of context and world knowledge. Jenkins et al. (2003) have shown that the mean reading rate of fourth graders with dyslexia was uniformly discrepant from skilled readers both in context and list. However, children with dyslexia seemed to benefit less than skilled readers from the context: their reading rate in text was 1.19 times of the rate of list reading, whereas in skilled readers the figure was 1.67 (Jenkins et al., 2003).

According to the verbal efficiency theory (Perfetti, 1985), deficiencies in children's word reading proficiency affect their fluency skills. A certain level of word reading proficiency seems to be needed before cognitive resources may be released for the language processing needed in fluent text reading (Kim et al., 2012).

Skillful readers can identify the meaning of familiar words rapidly just by sight without effort (Ehri, 2005). Other factors besides activation of lexical representations may also speed up word recognition in text reading. Stanovich (1980) found that context allows readers to anticipate possible upcoming words, while eye movement studies have shown that it is possible to get information of the next word parafoveally before fixating on it (Hyönä, 2011). Barker, Torgesen, and Wagner (1992) demonstrated that orthographic skills have a much stronger influence on reading speed of text, compared with the speed of single word identification: 20% vs. 5%, respectively. Deficiency in fluent access to word representations (i.e., poor orthographic skills) would therefore affect more reading speed of text in context and thus reduce the difference in reading speed between text and single words. Longitudinal design, such as the one used in our study, can reveal whether and at what age children with familial risk and dyslexia acquire sufficient word decoding skills for the release of cognitive recourses in language processing in order to speed up reading text in context compared with word list reading.

The Present Study

In summary, our study addresses three questions. First, what is the stability of reading and spelling skills after the early reading acquisition phase? Second, what is the effect of familial risk on reading and spelling development? We compared the development of reading speed, reading accuracy, and spelling across Grades 2, 3, and 8 in three groups of children: (a) those with both dyslexia and familial risk, (b) those without dyslexia but with familial risk, and (c) control children with no dyslexia and without familial risk. Third, are reading speed differences in varying reading tasks and materials (word list, text and, pseudoword text) similar across the three groups of participants and across Grades 2, 3, and 8?

Method

Participants

All children ($N = 173$) in this study were participants of the Jyväskylä Longitudinal Study of Dyslexia (JLD; e.g., Lyytinen et al., 2008). They were originally selected for one of two groups: with familial risk for dyslexia or without familial risk for dyslexia.¹ For this study, children were further allocated to three groups according to their reading and spelling skills at the end of Grade 2 and familial risk status: (a) children with dyslexia and familial risk (Dys_FR, $n = 35$), (b) children with no dyslexia and with familial risk (NoDys_FR, $n = 66$), and (c) a control group of children with no dyslexia and without familial risk (C, $n = 72$).² (See later descriptions of the familial risk and dyslexia). Characteristics of the groups are presented in Table 1. There were no differences between the groups in the parents' age or education or in the children's performance IQ, age, or gender distribution. However, the verbal IQ in the Dys_FR group was lower than in the NoDys_FR and C groups, $F(2, 169) = 6.63, p < .01$.

All the children spoke Finnish as their native language and had no mental, physical, or sensory impairments. An exclusion criterion was both verbal (VIQ) and performance IQ (PIQ) being below 80, which was assessed in Grade 2 using the Wechsler Intelligence Scale for Children (3rd ed.; WISC—III;

Wechsler, 1991). Four performance scale subtests (Picture Completion, Block Design, Object Assembly, and Coding) and five verbal scale subtests (Similarities, Vocabulary, Comprehension, Series of Numbers, and Arithmetic) were used to estimate the PIQ and VIQ, respectively. None of the participants were excluded according to the exclusion criterion. All participants attended regular classroom education.

Familial Risk: Screening of the Families

The children were originally selected from among 9,368 newborns born in the province of Central Finland between April 1993 and July 1996. The selection was made using a three-stage procedure: (a) a short parental questionnaire including three questions concerning difficulties in learning to read and spell among parents and their close relatives (8,417 respondents); (b) a detailed parental questionnaire concerning the reading history, the persistence of reading and spelling difficulties, and the reading habits of parents and their close relatives (3,130 respondents); and (c) testing of the reading and spelling skills (410 parents).

For the child to be originally included in the familial risk group ($n = 108$), either of the parents had to show deficient performance in oral text reading or spelling and in single word reading tasks tapping phonological and orthographic processing. In addition, a reported onset of literacy problems during early school years and a first-degree relative with corresponding difficulties were required for inclusion in the familial risk group. In the group without familial risk, both parents ($n = 92$) had no reported family history for dyslexia and had a z score above -1.0 in all reading and spelling tasks described previously. The IQ of all parents, assessed with the Raven B, C, and D matrices (Raven, Court, & Raven, 1992), had to be equal to or above 80 (for full details of recruitment, see Leinonen et al., 2001).

Identification of Children With Dyslexia in Grade 2

The identification of dyslexia was based on performance in five tasks (descriptions of the tasks will follow): (a) oral word and pseudoword reading, (b) oral text reading, (c) oral pseudoword text reading, (d) oral word list reading, and (e) spelling words and pseudowords. Four measures of reading speed were calculated: (a) mean response time (reaction time + response duration) of correctly read words and pseudowords presented one by one, (b) the number of read words per minute in oral text reading task, (c) the number of pseudowords read per minute in oral pseudoword text reading, and (d) the number of correctly read words in 2 min in oral word list reading. Respectively, four measures of reading and spelling accuracy were calculated: the number of (a) correctly read words and pseudowords presented one at a time, (b) correctly read words in oral text reading, (c)

¹ From the 200 children originally screened, 18 children refused to take part in the Grade 8 assessments, of whom three were from the group of children with reading disability and familial risk (Dys_FR), four from the group of children with no reading disability and with familial risk (NoDys_FR), and 11 were from the control group (children with no reading disability and without familial risk).

² Nine children without familial risk fulfilled the criteria for reading disability at the end of Grade 2 and were excluded from this study as in other studies examining the continuity of the genetic risk.

Table 1

Characteristics of Parents and Their Children in the Three Groups: Children With Dyslexia and Familial Risk, Children With No Dyslexia but With Familial Risk, and Control Children With No Dyslexia and Without Familial Risk

Variable	Dys_FR			NoDys_FR			Controls			Paired group comparisons
	M	SD	N (35)	M	SD	N (66)	M	SD	N (72)	
Parents										
Mother										
Age	29.62	4.26		29.32	4.22		29.67	4.10		Dys_FR = NoDys_FR = C
Education	4.09	1.42		4.40	1.44		4.60	1.34		
Father										
Age	31.53	5.36		31.64	5.04		32.75	5.34		
Education	3.61	0.99		3.71	1.41		3.75	1.48		
Children										
WISC-III										
Verbal IQ	94.17	9.75		100.85	11.77		102.38	11.10		Dys_FR < NoDys_FR* Dys_FR < C**
Performance IQ	97.26	14.25		100.77	11.79		103.23	14.10		
Age (years)										
Grade 2	8.98	0.34		8.99	0.32		8.98	0.29		Dys_FR = NoDys_FR = C
Grade 3	9.99	0.45		9.85	0.32		9.83	0.29		
Grade 8	14.48	0.44		14.30	0.54		14.35	0.28		
Gender			19 girls, 16 boys			32 girls, 34 boys			34 girls, 38 boys	

Note. Groups: Dys_FR = Dyslexia with familial risk; NoDys_FR = No dyslexia with familial risk; and Controls = Control children with no dyslexia and without familial risk. Parental education was classified using a 7-point scale: 1 = *only comprehensive school (CS)*; 2 = *CS and short-term vocational courses*; 3 = *CS and vocational school degree*; 4 = *CS and vocational college degree*; 5 = *CS and lower university degree/polytechnic degree*; 6 = *upper secondary general school and lower university degree/polytechnic degree*; 7 = *CS or upper secondary general school and higher university degree (master's or doctorate)*. WISC-III = Wechsler Intelligence Scale for Children (3rd ed.).

* $p \leq .05$. ** $p \leq .01$.

correctly read pseudowords in oral pseudoword text reading, and (d) correctly written words and pseudowords, presented one by one in a dictation task.

For the identification of dyslexia, a two-step procedure was used. First, a cutoff criterion for deficient performance was defined for each of the eight measures using the 10th percentile of the control group's performance. Second, a child was considered to have dyslexia if she or he scored below the criteria in at least three out of four measures of reading speed or in at least three out of four measures in reading and spelling accuracy. In addition, a child who scored below the criteria both in two speed and two accuracy measures was considered to have dyslexia.

Measures

Trained testers assessed reading and spelling skills individually in a laboratory setting with four different tasks in Grade 2 (June), Grade 3 (April), and Grade 8 (November) as a part of the JLD assessment procedure: (a) oral text reading, (b) oral pseudoword text reading, (c) oral word list reading, and (d) spelling pseudowords. In all reading tasks, children were instructed to read "as quickly and accurately" as they could. Two different measures were calculated from each task: reading speed (the number of letters read in 1 s) and reading accuracy (the percentage of correctly read items). Arithmetical means, calculated from the three oral reading tasks described previously, were used as composite measures of reading speed and reading accuracy separately for Grades 2, 3, and 8. The Cronbach's alpha reliability for the reading speed composite was .93, .89, and .88 and for the reading accuracy composite .82, .83, and .75, in Grades 2, 3, and 8, respectively.

Oral text reading (Grades 2, 3, and 8). At each grade level, participants read aloud an age-appropriate text for oral text reading. In Grade 2, the text (title "Exciting Journeys") consisted of 19 sentences in five paragraphs with a total of 124 words/877 letters (mean word length = 7.07 letters, and mean sentence length = 6.53 words). For Grade 3, the text (title "Useless Belongings") consisted of 18 sentences in four paragraphs and a total of 189 words/1,154 letters (mean word length = 6.11 letters, and mean sentence length = 10.50 words). Finally, the Grade 8 text (title "Fjelds of Lapland") consisted of 16 sentences in three paragraphs and a total of 207 words/1,591 letters (mean word length = 7.68 letters/word, and mean sentence length = 12.94 words). Reading performance was recorded on a tape recorder (Grades 2 and 3) or a laptop computer (Grade 8). The total time to read the text was measured with a stop watch. The tapes and sound files were subsequently used to check the scoring of the children's accuracy and speed. To assess the reliability of accuracy scoring, two trained coders independently scored accuracy in a randomly selected 10% of the sample, and the interrater agreement was .98.

Oral pseudoword text reading (Grades 2, 3, and 8). Participants read aloud a short text made up of 19 pseudowords/137 letters (Grade 2) or 38 pseudowords/277 letters (Grades 3 and 8). The words and structure of the sentences resembled real Finnish in form but had no meaning. The mean word length was 7.21 letters/word in Grade 2 and 7.29 letters/word in Grades 3 and 8. Similarly to the oral text reading, the child's reading performance was recorded, and correctness of reading and time spent on reading were checked. In 10% of the sample, each pseudoword was judged by two coders as correctly or incorrectly read, and the interrater agreement was .95.

Oral word list reading (Grades 2, 3, and 8). In the Lukilasse standardized reading test (Häyrynen, Serenius-Sirve, & Korkman, 1999), the participant has 2 min to read aloud as many words as possible from a 90-item (Grade 2) or 105-item (Grade 3) list, assembled vertically in columns. The same list that was used in Grade 3 was administered also in Grade 8, but the time limit was reduced to 1 min. The length of the words increased gradually, ranging from three to 18 letters/word in Grade 2 and from three to 22 letters/word in Grades 3 and 8. The mean length of the words was 9.08 letters in Grade 2 and 9.57 letters in Grades 3 and 8. A trained tester marked the incorrectly read words as the child was reading aloud. The correctness of tester markings was checked by another listener in 10% of the sample using the recordings, and the interrater reliability was .99.

Oral word and pseudoword reading (used only for the identification procedure of dyslexia in Grade 2). Children read aloud three- and four-syllable words and pseudowords (10 of each type, altogether 40 items) presented one by one with the program Cognitive Workshop (Seymour, 1995) on a computer screen.

Spelling pseudowords (Grades 2, 3, and 8). We measured spelling accuracy with a list of pseudowords consisting of 12 four-syllable items in Grades 2 and 3 and 20 three- to five-syllable items in Grade 8. Participants listened through headphones as a computer presented the items twice with a 2-s interval. Each pseudoword was scored as correct if all the phonemes were correctly written without missing or extra letters. The percentage of correctly written pseudowords was used as the spelling accuracy measure separately for each grade. Cronbach's alpha reliability coefficients were .80, .71, and .70 for Grades 2, 3, and 8, respectively.

Spelling words and pseudowords (used only for the identification procedure of dyslexia in Grade 2). Participants used a pencil to write 6 four-syllable words and 12 four-syllable pseudowords presented similarly as described previously. Each stimulus (word or pseudoword) was scored as correct if the participants wrote all the phonemes correctly without missing or extra letters. The percentage of correctly written words/pseudowords was used

as the spelling accuracy measure. Cronbach's alpha reliability coefficient was .87.

Results

Distributions and Stability of Literacy Skills

All distributions of reading speed measures were normal or close to normal. The distributions of reading and spelling accuracy, instead, showed a ceiling effect in all tasks in all grades. The ceiling effect was particularly clear in oral word list reading accuracy, with 82.5%, 89.0%, and 98.3% of the participants exceeding 90% accuracy in Grades 2, 3, and 8, respectively. The ceiling effect also appeared in oral text reading accuracy, where the portion of children above the 90% accuracy level was 79.2%, 86.2%, and 89.5% in Grades 2, 3, and 8, respectively. We applied logarithmic transformation to correct the distribution in the oral text reading task, whereas the distributions of the tasks for oral word list reading and spelling pseudowords could not be normalized. Because of the nonnormal distributions, we conducted both parametric and nonparametric analyses when applicable. As all conclusions derived from the parametric and nonparametric analysis results were identical, we report only the parametric results. In reading and spelling accuracy measures, one to four extreme outliers were moved to the tail of the distribution before analyses to avoid overemphasizing their effects on results. No participants were dropped from the sample.

Table 2 presents correlations between overall (averaged composite measure of) reading speed and accuracy as well as spelling accuracy. For the reading speed measures, the correlations between performance across different grades were high (.72–.88). For reading accuracy measures, the correlations varied from moderate to high (.51–.69), and for spelling measure they were moderate (.41–.59).

Table 2
Spearman Correlations of Reading Speed, Reading Accuracy, and Spelling Accuracy in Grades 2, 3, and 8

Variable	Reading speed			Reading accuracy			Spelling accuracy	
	1	2	3	4	5	6	7	8
Reading speed								
1. Grade 2	—							
2. Grade 3	.88***	—						
3. Grade 8	.72***	.78***	—					
Reading accuracy								
4. Grade 2	.53***	.55***	.50***	—				
5. Grade 3	.63***	.55***	.54***	.69***	—			
6. Grade 8	.47***	.47***	.38***	.51***	.62***	—		
Spelling accuracy								
7. Grade 2	.49***	.46***	.41***	.52***	.49***	.40***	—	
8. Grade 3	.40***	.40***	.32***	.52***	.49***	.40***	.59***	—
9. Grade 8	.36***	.35***	.30***	.37***	.39***	.31***	.41***	.44***

Note. $N = 173$ in all correlations between the Grade 3 and Grade 8 measures, and $N = 171$ in correlations where a Grade 2 measure is included. Reading speed and accuracy are the arithmetic means of three oral reading tasks at each grade: word list, text, and pseudoword text. Spelling accuracy is the percentage of correctly spelled items in pseudoword spelling task.

*** $p < .001$.

Continuity of the Familial Risk: Group Differences in the Development of Literacy Skills

We examined the development of reading speed and accuracy as well as spelling in the groups with mixed-design analyses of variance (ANOVAs) including grade (2, 3, and 8) as the within-subject factor and group (Dys_FR, NoDys_FR, and controls) as the between-subjects factor. For both reading speed and accuracy, a composite score was used as the measure at each grade level (arithmetic mean from the three tasks: list, text and pseudoword text reading). Figure 1 presents the development of each skill in the three groups. To evaluate the gain children made between two grades (Grade 2 and Grade 3; Grade 3 and Grade 8), a difference score was calculated by subtracting the corresponding means from each other. We used one-way ANOVAs to study group differences in these gains as well as in separate tasks of reading speed and accuracy and spelling in each grade. In the post hoc pairwise comparisons, we used either Bonferroni (when equal variances) or Dunnett's T3 (when unequal variances) correction when evaluating the significances of group differences (see Table 3).

In the mixed-design ANOVA for the reading speed composite, both main effects, grade and group, were significant, $F(1.62, 271.69) = 724.69, p < .001, \eta_p^2 = .81$, and $F(2, 168) = 49.79, p < .001, \eta_p^2 = .37$, respectively, as was the Grade \times Group interaction, $F(3.23, 271.69) = 2.93, p < .05, \eta_p^2 = .03$. For further evaluating the dissimilarity between the groups in the development of reading speed between Grade 2 and Grade 3 as well as between Grade 3 and Grade 8, the tests of within-subject contrast for the Grade \times Group interaction were used. The effect was significant for the development between Grade 2 and Grade 3, $F(2, 168) = 7.31, p < .001, \eta_p^2 = .08$, but not for the development between Grade 3 and Grade 8, which suggests that the reading speed development differed between groups in Grade 2 and Grade 3, but not in Grade 3 and Grade 8. The ANOVA post hoc pairwise

comparisons (with Bonferroni corrections for significance) of the reading speed improvement between Grade 2 and Grade 3 showed that children in the Dys_FR group improved their overall reading speed more than the children in the control group ($p < .001$) and the NoDys_FR ($p < .01$) group between Grade 2 and Grade 3. However, the children in the Dys_FR group still did not reach the level of the other two groups as shown by the post hoc ANOVA comparisons at Grade 3 (see Table 3). The overall reading speed of children in the Dys_FR group was about 50%, 65%, and 75% in Grades 2, 3, and 8, respectively, from the reading speed of children in the two other groups (NoDys_FR and controls). In Grade 8, the overall reading speed of Dys_FR children was approximately at the level of third graders compared with the two other groups, indicating a lag of 5 years in development. Effect sizes were estimated (Cohen's d computed using pooled standard deviation), and they were large not only for Grades 2 and 3 but also for Grade 8: Dys_FR versus NoDys_FR ($d = 1.21$) and Dys_FR versus control group ($d = 1.73$). No significant differences between the NoDys_FR and control group were found in ANOVA post hoc pairwise comparisons (with Bonferroni corrections for significance) of the gain children made in overall reading speed between Grade 2 and Grade 3 or between Grade 3 and Grade 8.

For each task in each grade level, we separately conducted one-way ANOVAs. These showed that children in the Dys_FR group read slower in all tasks throughout Grades 2, 3, and 8 than the two other groups (see Table 3). The two groups without dyslexia (NoDys_FR and controls) did not differ from each other in any of the reading speed measures, although the effect sizes varied from small to moderate (.15–.42).

In the analysis of the reading accuracy composite, both main effects, grade and group, were significant, $F(1.75, 293.92) = 104.28, p < .001, \eta_p^2 = .38$, and $F(2, 168) = 83.12, p < .001, \eta_p^2 = .25$,

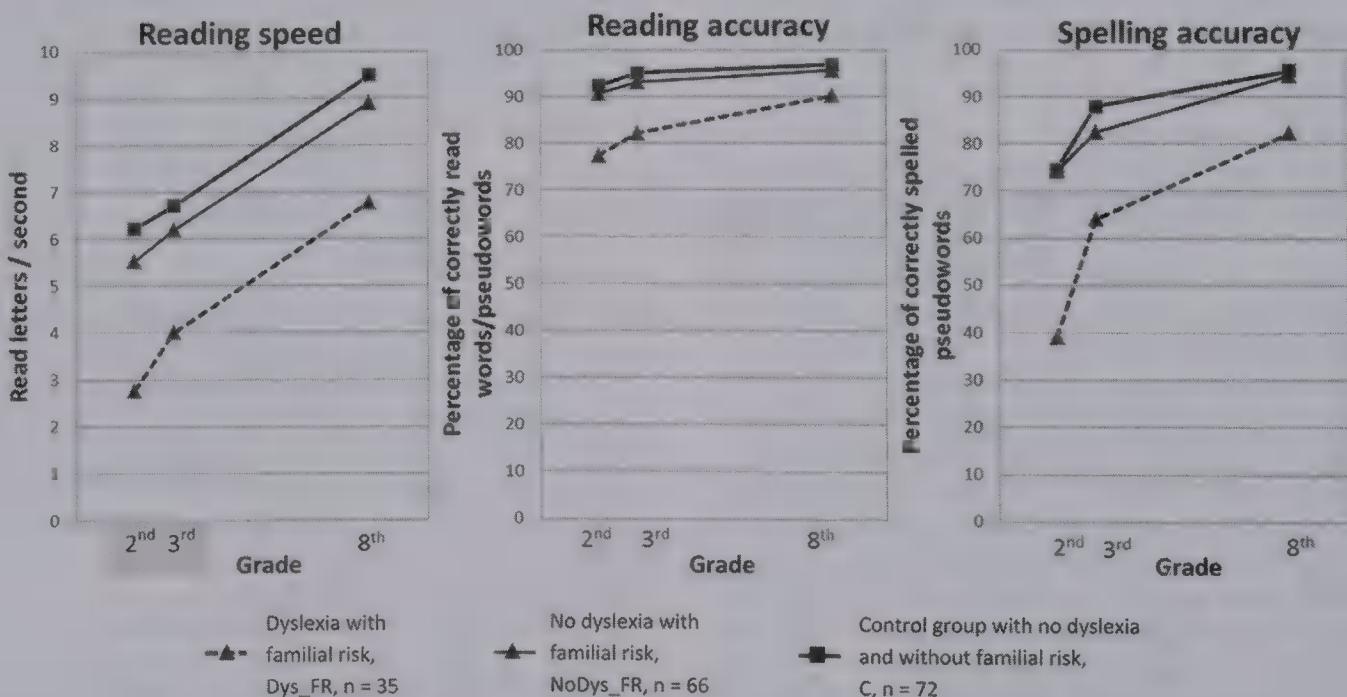


Figure 1. Reading speed and accuracy (composite means) and pseudoword spelling accuracy in the three groups: children with dyslexia and familial risk, children with no dyslexia but with familial risk, and control children at Grades 2, 3, and 8.

Table 3

Descriptive Statistics and Group Comparisons of Dys_FR, NoDys_FR and Control Groups With One-Way Analyses of Variance

Variable	Dys_FR (N = 35)		NoDys_FR (N = 66)		C (N = 72)		F ^a	Effect size		
	M	SD	M	SD	M	SD		Dys_FR vs. NoDys_FR	Dys_FR vs. C	NoDys_FR vs. C
Reading speed										
Overall										
Grade 2	2.77 _x	0.83	5.53 _y	1.55	6.22 _y	1.88	56.81***	2.07	2.12	.40
Grade 3	4.00 _x	1.07	6.21 _y	1.63	6.71 _y	1.72	36.05***	1.53	1.76	.30
Grade 8	6.78 _x	1.74	8.96 _y	1.87	9.49 _y	1.51	30.79***	1.21	1.73	.32
Word list										
Grade 2	2.20 _x	0.72	4.87 _y	1.77	5.54 _y	2.09	43.27***	1.80	1.88	.34
Grade 3	3.61 _x	1.13	6.34 _y	2.01	6.88 _y	1.97	38.41***	1.57	1.87	.27
Grade 8	6.96 _x	2.02	9.29 _y	2.61	9.66 _y	2.30	16.23***	0.97	1.23	.15
Text										
Grade 2	4.01 _x	1.51	7.92 _y	2.37	8.89 _y	2.55	54.01***	1.87	2.15	.39
Grade 3	4.92 _x	1.61	8.11 _y	2.17	8.81 _y	2.24	41.99***	1.61	1.89	.32
Grade 8	8.59 _x	1.91	11.17 _y	2.06	11.81 _y	1.58	37.13***	1.30	1.93	.35
Pseudoword text										
Grade 2	2.09 _x	0.75	3.80 _y	1.06	4.22 _y	1.43	39.52***	1.79	1.70	.33
Grade 3	3.47 _x	1.19	4.17 _y	1.25	4.43 _y	1.41	6.34***	0.57	0.72	.19
Grade 8	4.54 _x	1.42	6.37 _y	1.55	6.94 _y	1.69	26.08***	1.23	1.50	.35
Reading accuracy										
Overall										
Grade 2	77.38 _x	11.49	90.62 _y	6.76	92.30 _y	4.63	53.22***	1.54	2.04	.30
Grade 3	82.21 _x	7.31	93.15 _y	5.71	95.05 _y	3.88	69.73***	1.75	2.51	.40
Grade 8	90.14 _x	6.09	95.70 _y	4.46	96.76 _y	2.39	31.12***	1.11	1.72	.31
Word list										
Grade 2	87.58 _x	8.97	94.72 _y	4.38	96.28 _y	3.27	32.91***	1.46	1.92	.41
Grade 3	90.84 _x	6.71	96.04 _y	3.77	97.35 _y	2.87	28.67***	1.05	1.50	.40
Grade 8	97.34 _x	3.66	99.54 _y	1.11	99.41 _y	1.85	13.84***	0.96	0.83	.08
Text										
Grade 2	85.25 _x	10.05	94.52 _y	4.02	94.64 _y	4.69	33.24***	1.39	1.40	.03
Grade 3	90.00 _x	5.52	95.32 _y	3.93	97.27 _z	1.89	44.95***	1.18	2.15	.66
Grade 8	90.35 _x	5.83	95.57 _y	4.16	96.78 _y	2.12	31.84***	1.10	1.78	.38
Pseudoword text										
Grade 2	59.10 _x	21.82	82.63 _y	15.72	85.94 _y	10.52	38.10***	1.32	1.83	.25
Grade 3	65.52 _x	15.05	87.86 _y	12.15	90.34 _y	8.38	58.58***	1.71	2.32	.24
Grade 8	82.14 _x	12.94	92.03 _y	9.63	94.03 _y	5.21	20.84***	0.92	1.44	.27
Spelling accuracy										
Grade 2	39.05 _x	28.99	74.36 _y	17.32	74.53 _y	19.10	39.96***	1.62	1.59	.01
Grade 3	64.05 _x	21.18	82.45 _y	16.67	87.96 _y	14.19	24.44***	1.01	1.45	.36
Grade 8	82.29 _x	15.36	94.32 _y	7.23	95.49 _y	4.45	29.90***	1.13	1.45	.20

Note. Groups: Dys_FR = Dyslexia with familial risk; NoDys_FR = No dyslexia with familial risk; and C = Control children with no dyslexia and without familial risk. Groups with different subscript letter (x, y, or z) were significantly different in the post hoc pair-wise comparisons of analyses of variance F tests ($p < .05$). Bonferroni or Dunnett's T3 corrections were used, depending on equality or inequality of the variances. Effect sizes were estimated with Cohen's d (computed with pooled standard deviations). Overall reading speed = arithmetic mean of the number of read letters/second in word list, text, and pseudoword text reading. Overall reading accuracy = arithmetic mean of the percentage of correctly read words/pseudowords in word list, text, and pseudoword text reading. Spelling accuracy = percentage of correctly written pseudowords.

^a Degrees of freedom varied between 2,165 and 2,170 due to missing data in single measures.

*** $p \leq .001$.

.50, respectively) as well as, the Grade \times Group interaction, $F(3.50, 293.92) = 11.72, p < .001, \eta^2 = .12$. The test of within-subject contrasts for the Grade \times Group interaction was not significant between Grade 2 and Grade 3, but it was significant between Grade 3 and Grade 8, $F(2, 168) = 18.78, p < .001, \eta^2 = .18$, a result that suggests that there was a difference in the developmental pace of reading accuracy between the groups in Grade 3 and Grade 8. The ANOVA post hoc pairwise comparisons (with Dunnett's T3 corrections for significance) showed that between Grade 3 and Grade 8, the children in the Dys_FR group developed faster in reading accuracy than did the children in the other two groups (both $p < .001$). However, as with reading speed

described earlier, the children in the Dys_FR group did not quite reach the level of the other two groups (see Figure 1 and Table 3). In the Dys_FR group, the overall reading accuracy level reached 90% in Grade 8, whereas the two groups without dyslexia (NoDys_FR and controls), had reached the 90% level in overall reading accuracy already at the end of Grade 2. Effect sizes were large not only in Grades 2 and 3 but also for the Grade 8 group comparisons in reading accuracy: Dys_FR versus NoDys_FR ($d = 1.11$) and Dys_FR versus Control group ($d = 1.72$). No significant differences between the NoDys_FR and control group were found in ANOVA post hoc pairwise comparisons (with Bonferroni corrections for significance) of the gain children made in overall

reading accuracy between Grade 2 and Grade 3 or between Grade 3 and Grade 8.

One-way ANOVAs, done separately for each task in each grade, showed that children in the Dys_FR group made more errors in all reading tasks throughout Grades 2, 3, and 8 than the two other groups (see Table 3). The two groups without dyslexia did not differ from each other in any of the reading accuracy measures, except in text reading accuracy in Grade 3. Effect sizes were small or medium (.03–.66) between these two groups in reading accuracy measures throughout Grades 2, 3, and 8.

In the analysis of pseudoword spelling both main effects, grade and group, were significant, $F(1.87, 314.79) = 181.98, p < .001$, $\eta_p^2 = .52$, and $F(2, 168) = 49.57, p < .001, \eta_p^2 = .37$, respectively. Also the Grade \times Group interaction was significant, $F(3.75, 314.79) = 11.86, p < .001, \eta_p^2 = .12$. The test of within-subject contrasts for the Grade \times Group interaction was significant between Grade 2 and Grade 3, as well as between Grade 3 and Grade 8, $F(2, 168) = 8.64, p < .001, \eta_p^2 = .09$, and $F(2, 168) = 5.84, p < .01, \eta_p^2 = .06$, respectively. The ANOVA post hoc pairwise comparisons (with Dunnett's T3 corrections for significance) showed that between Grade 2 and Grade 3, children in the Dys_FR group improved their spelling accuracy more than children in the control group ($p < .05$) and the NoDys_FR group ($p < .01$), and more than the control group ($p < .01$) between Grade 3 and Grade 8. Note, however, that the starting point of spelling accuracy in the two groups without dyslexia was approximately twice as high as in the Dys_FR group (with accuracy percentages of 73% vs. 39%). Although children in the Dys_FR group made better progress in spelling accuracy than the NoDys_FR and control groups, they reached accuracy level of 82.39% in Grade 8, which is comparable to the level of third graders in the other two groups (see Table 3). The group differences in Grade 8 were confirmed by effect sizes, which were large: 1.13 (Dys_FR vs. NoDys_FR) and 1.45

(Dys_FR vs. control group). The two groups without dyslexia, NoDys_FR and controls, reached close to 95% accuracy level in pseudoword spelling in Grade 8 and did not differ from each other in any of the spelling measures. Effect sizes were small or moderate (.01–.36).

Differences in Reading Speed According to Task in Different Groups

To see whether the differences in reading speed between the three tasks were similar in the three groups, we performed three separate mixed-design ANOVAs. Task (text vs. pseudoword text, text vs. word list, or pseudoword text vs. word list) was used as the within-subject factor, and group (Dys_FR, NoDys_FR, and control group) as the between-subjects factor. We did all ANOVAs separately for each grade level (2, 3, and 8) to see whether the differences between tasks were similar in each grade. Because nine mixed-design ANOVAs were conducted, stricter than usual significance cutoffs were used to avoid family-wise errors. This was done by dividing the commonly used significance levels by the number of ANOVAs done. As a follow-up analysis, we compared performance in the different tasks within each group using paired sample t tests. Figure 2 presents group differences in the three reading speed tasks in Grades 2, 3, and 8. Table 4 presents F values and estimates of the effect sizes of the mixed-design ANOVAs.

Word list versus pseudoword text. We compared reading speed in word list and pseudoword text reading first to see the effect of lexicality. Both main effects, task and group, were significant. The interaction Task \times Group was significant in Grades 2 and 3. In paired sample t tests, the difference between tasks was not significant in the Dys_FR group, a result that suggests that children in this group read word lists and pseudoword texts at equal speeds. In the NoDys_FR group and in the control group,

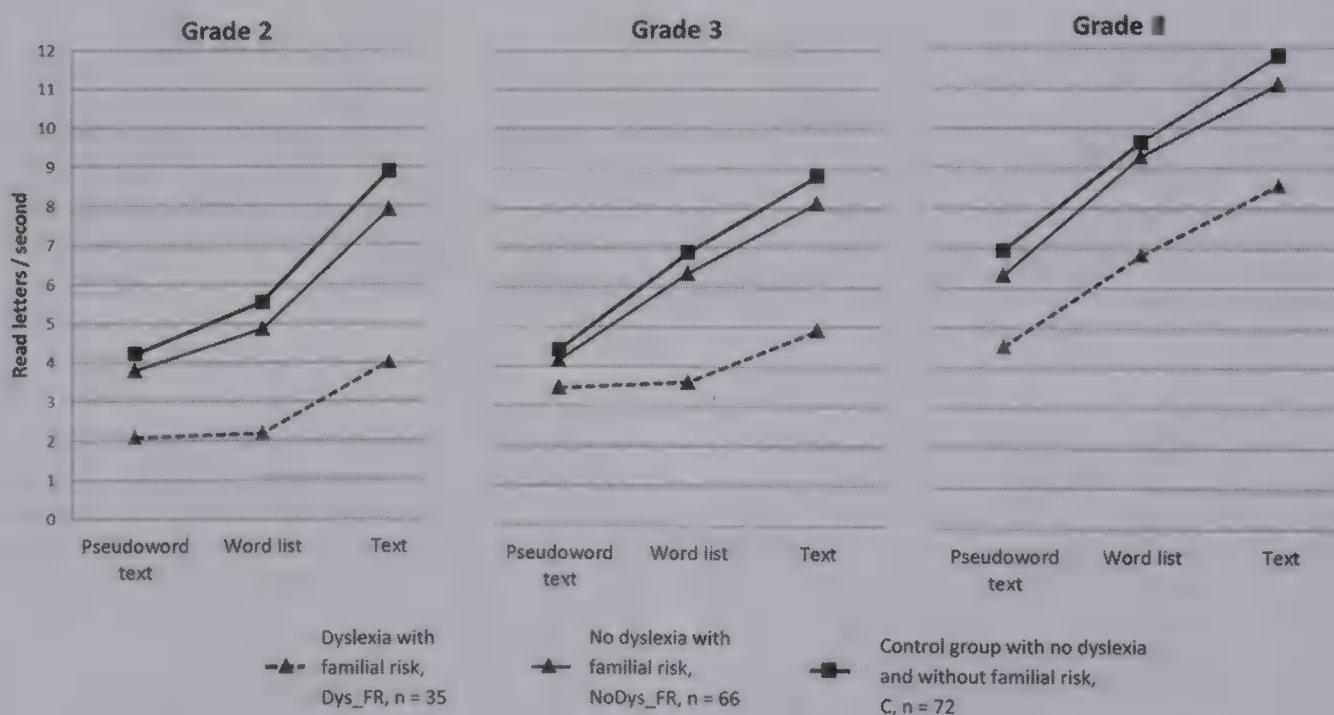


Figure 2. Reading speed means in pseudoword text, word list, and text reading tasks in the three groups: children with dyslexia and familial risk, children with no dyslexia but with familial risk, and control children at Grades 2, 3, and 8.

Table 4

F Values and Estimates of Effect Sizes From Mixed-Design Analyses of Variance With Reading Speed as the Dependent Measure, Task as the Within-Subjects Factor, and Group as the Between-Subjects Factor

Compared tasks	Main effect of task	Effect size ^a	Main effect of group	Effect size ^a	Interaction effect of Task * Group	Effect size ^a
Word list versus pseudoword text						
Grade 2	$F(1, 168) = 176.47^{***}$.31	$F(2, 168) = 47.54^{***}$.36	$F(2, 168) = 12.55^{***}$.13
Grade 3	$F(1, 170) = 215.07^{***}$.56	$F(2, 170) = 25.74^{***}$.23	$F(2, 170) = 36.76^{***}$.30
Grade 8	$F(1, 166) = 244.49^{***}$.60	$F(2, 166) = 26.77^{***}$.24	$F(2, 166) = 11.12$.01
Word list versus text						
Grade 2	$F(1, 168) = 597.92^{***}$.78	$F(2, 168) = 54.67^{***}$.39	$F(2, 168) = 14.55^{***}$.15
Grade 3	$F(1, 170) = 292.26^{***}$.63	$F(2, 170) = 44.36^{***}$.34	$F(2, 170) = 13.14$.04
Grade 8	$F(1, 169) = 152.17^{***}$.47	$F(2, 169) = 29.98^{***}$.26	$F(2, 169) = 11.06$.01
Text versus pseudoword text						
Grade 2	$F(1, 168) = 1591.75^{***}$.78	$F(2, 168) = 58.93^{***}$.41	$F(2, 168) = 27.09^{***}$.24
Grade 3	$F(1, 170) = 1549.10^{***}$.76	$F(2, 170) = 30.92^{***}$.27	$F(2, 170) = 35.36^{***}$.29
Grade 8	$F(1, 167) = 2249.39^{***}$.93	$F(2, 167) = 34.66^{***}$.29	$F(2, 167) = 16.12^*$.07

Note. Groups: Dys_FR = Dyslexia with familial risk, $n = 35$; NoDys_FR = No dyslexia with familial risk, $n = 66$; and C = Control children with no dyslexia and without familial risk, $n = 72$.

^a Effect size = partial eta square.

Because multiple mixed-design analyses of variance were conducted, stricter-than-usual cutoffs for significance were used: * $p \leq .005$. *** $p \leq .0001$.

children read the word lists about 1–2 letters/second faster than pseudoword texts, $t(64) = 6.47$, $p < .001$, and $t(65) = 12.14$, $p < .001$, in Grade 2 and Grade 3, respectively, for the NoDys_FR group, and $t(70) = 8.87$, $p < .001$, and $t(71) = 15.13$, $p < .001$, for the control group in Grade 2 and Grade 3, respectively). In Grade 8, all groups read word lists about 2.5 letters/second faster than pseudoword texts, $t(32) = 11.36$, $t(64) = 11.40$, and $t(70) = 9.60$, all $p < .001$, for the Dys_FR, NoDys_FR, and control group, respectively (see Table 3).

Word list versus text. We compared reading speed in word lists and text reading to see the effect of context on reading speed. Both main effects, task and group, were significant in all grades (2, 3, and 8). The interaction Task \times Group was significant only in Grade 2. In paired sample t tests, the difference between tasks was significant in all groups, $t(34) = 9.01$, $t(64) = 17.03$, and $t(70) = 19.61$, all $p < .001$, for the Dys_FR, NoDys_FR, and control group, respectively, indicating that all groups read words in context faster than isolated words. However, the ANOVA post hoc pairwise group comparisons (with Bonferroni corrections for the significance) indicated that the difference in reading speed between word lists and texts was smaller in the Dys_FR group than in the NoDys_FR group and in the control group (both $p < .001$). In Grade 3 and in Grade 8, all groups read text faster than they read word lists, $t(34) = 6.92$ and $t(34) = 7.48$; $t(64) = 10.51$ and $t(64) = 8.26$; and $t(71) = 14.62$ and $t(71) = 8.30$, all $p < .001$, for the Dys_FR, NoDys_FR, and control group in Grade 3 and in Grade 8, respectively.

Text versus pseudoword text. Finally, we compared reading speed between text and pseudoword text reading tasks to see the effect of the lexicality and meaning of the text. Both main effects, task and group, as well as the interaction Task \times Group in Grades 2, 3, and 8 were significant. In the ANOVA post hoc pairwise group comparisons (with Bonferroni corrections for the significance), the difference in reading speed between texts and pseudoword texts was smaller in the Dys_FR group than in the NoDys_FR and control groups (all $p < .001$ in Grades 2 and 3, and both $p < .01$ in Grade 8; 2 vs. 4 letters/second, respectively, in Grades 2 and 3, and 4 vs. 5 letters/second in Grade 8; see Table 3).

Discussion

In this study, we examined three aspects of literacy development: the stability of literacy skills after the initial reading acquisition phase across Grades 2, 3, and 8; the effect of familial risk on literacy skill development during this period; and the effects of different types of reading material (word list, text, and pseudoword text) on reading speed. We compared the development of three groups: children with familial risk and dyslexia (the Dys_FR group), children with familial risk but without dyslexia (the NoDys_FR group), and a control group of children with no dyslexia and without familial risk.

We found high stability for reading speed development, whereas in reading and spelling accuracy, the development was moderately stable from the second to the eighth grade. Children with familial risk and dyslexia (the Dys_FR group) did not catch up to the other two groups in reading speed, reading accuracy, or spelling, although they progressed more than the other two groups in reading speed between Grade 2 and Grade 3, in reading accuracy between Grade 3 and Grade 8, and in spelling accuracy throughout the follow-up. The Dys_FR group's literacy skills in Grade 8 were overall comparable to the level of the third graders in the two other groups. The children with familial risk but no dyslexia (NoDys_FR) did not differ significantly from the control group children in any of the assessed reading and spelling measures, except in text reading accuracy in Grade 3, although the effect sizes were often of moderate size between the two groups. The reading speed in children with familial risk and dyslexia varied less according to the type of reading material than in the two other groups in Grades 2 and 3, but this effect diminished in Grade 8.

In reading speed, the correlations across groups between the grades were high (.72–.88). This indicates high stability of development and is in line with earlier findings in consistent orthographies (Landerl & Wimmer, 2008; Parrila et al., 2005; Torppa et al., 2007). The size of the correlation between the assessments in Grades 2 and 8 was .72, which showed that even after 6 years of school attendance, the relative positions of individuals remained very similar. The nearly parallel developmental paths of the three

groups confirm the idea of stability in reading speed. The only exception, the faster progress made by children in the Dys_FR group in reading speed between Grades 2 and 3, could be interpreted to be a delayed developmental spurt that was made by normally developing children before the end of Grade 2. In previous Finnish studies of reading accuracy (Aunola, Leskinen, Onatsu-Arvilommi, & Nurmi, 2002; U. Leppänen, Niemi, Aunola, & Nurmi, 2004) as well as in a Finnish study of oral reading fluency (Parrila et al., 2005), initial reading level has been found to be negatively associated with the development of reading skill during the first two grades at school. Our finding that children in the NoDys_FR and control groups made only a little progress in reading speed between Grades 2 and 3 suggests that this kind of negative association between the initial level and further growth in reading speed continues to be true until the end of Grade 3. Between Grades 3 and 8, the development in the three groups was highly parallel, and we found no evidence suggesting either catching up or falling behind in any group. This supports the idea that differences between the groups are long-lasting, as has been found to be the case in Dutch (Boets et al., 2010; de Jong & van der Leij, 2003; van Bergen et al., 2011) and in English (Francis et al., 1996; Snowling et al., 2007) readers with and without dyslexia.

The consistent lag of the Dys_FR group in reading speed, present already at the beginning of the follow-up, could be expected because in transparent orthographies the main characteristic of dyslexia has been shown to be slow reading (e.g., de Jong & van der Leij, 2003; Landerl & Wimmer, 2008; Landerl, Wimmer, & Frith, 1997; Wimmer, 1996; Zoccolotti et al., 1999). The magnitude of the lag in Grade 8, approximately 5 years, was, however, larger than expected. De Jong and van der Leij (2003) have previously reported that Dutch children diagnosed with dyslexia in Grade 3 on the basis of reading fluency showed a delay of 3.5 years by the end of Grade 6 compared with normal readers in reading speed. In addition, in earlier studies that have used a reading-level matched group as controls, the age difference has usually been 3–4 years on average (Constantinidou & Stainthorp, 2009; Ziegler et al., 2003).

The stability in reading accuracy development was moderate to relatively high according to correlations (.51–.69) between Grades 2, 3, and 8 but lower than in reading speed. Correlations were somewhat lower than reported in previous studies of Finnish orthography (U. Leppänen et al., 2006; Parrila et al., 2005). The size of the correlations could be inflated by ceiling effects, but only for word list and text reading. In these tasks, where the items were real words, the percentage of correctly read words exceeded 90% before our first assessment point in this study (i.e., the end of Grade 2) in the NoDys_FR and control groups and in Grade 3 in the Dys_FR group. The accuracy percentages in the NoDys_FR and control groups are comparable to those reported earlier in transparent orthographies (Aro & Wimmer, 2003; de Jong & van der Leij, 2003). The ceiling effect also explains the finding that children in the Dys_FR group made better progress in reading accuracy between Grades 3 and 8 than children in the two other groups. After Grade 3, the children in the NoDys_FR and control groups simply had less room for development, having accuracy percentages at or above 96% in word list and text reading. In Grade 8, the mean percentage of correctly read words in the word list reading task was above 97% in all groups. Our finding that most children in the Dys_FR group also acquired accurate reading of

words is in line with the notion of de Jong and van der Leij (2003) that dyslexic children learning to read in a regular orthography eventually acquire sufficiently good skills in phonemic awareness to enable accurate decoding ability. However, reading accuracy concerning pseudoword items remained rather low in the Dys_FR group, even in Grade 8 (approximately 82%) and was equivalent to the accuracy of second graders in the NoDys_FR and control groups. This indicates persistent problems in phonological decoding among reading-disabled children when the demands of the task increases, in line with the findings of de Jong and van der Leij (2003). Previously, the parents of children with familial risk for dyslexia have been found to show difficulties in phonological decoding in the Jyväskylä Longitudinal Study of Dyslexia (P. H. T. Leinonen et al., 2001).

Problems in phonological decoding were seen especially clearly in pseudoword spelling, in which children in the Dys_FR group started the follow-up in Grade 2 with a very low accuracy percentage, 39%. Although they progressed faster than the other groups throughout the whole follow-up period, they remained behind children in the two other groups and ended up with a similar accuracy level as in pseudoword text reading (i.e., 82%) in Grade 8. This percentage is, however, much higher than the level of German-speaking Austrian children with reading and spelling difficulties: the mean of correctly spelled words for them was around 40%–45% in Grade 4 (Wimmer & Mayringer, 2002). This is probably due to the fact that in Germany a simple phoneme-grapheme translation is not sufficient for accurate spelling (Wimmer & Mayringer, 2002). In contrast to German, Finnish orthography has symmetrically transparent correspondences between phonemes and graphemes, that is, both from the point of view of reading and spelling. The stability in spelling was moderate (.41–.59) but somewhat lower than found earlier in Finnish (U. Leppänen et al., 2006). This discrepancy is probably due to this study's use of pseudoword items, whereas in U. Leppänen et al. (2006) a word-spelling task was used. In pseudoword spelling tasks, correlations have been found to be lower than between tasks including words (Lervåg & Hulme, 2010).

The greater gains in literacy skills by children in the Dys_FR group between Grades 2 and 3 could also be due to the extra support and intervention they have received. At Finnish schools, more than 20% of all school children in Grades 1–9 receive part-time special education at some point of the school year. This type of extra support is most frequent at the lowest grade levels, and the most common indication for part-time special education is problems in reading development. Altogether, 85.7% of children in the Dys_FR group received various kinds and amounts of extra support at school during Grades 1–3. This proportion is much bigger than the amounts of extra support in the NoDys_FR and control groups (34.8% and 11.1%, respectively). In addition, 48.6% of children in the Dys_FR group (4.5% in the NoDys_FR and none in the control group) took part in an intensive intervention study (55 hr within 14 weeks) organized by the JLD project, including speech and auditory training as well as practicing of reading and writing. However, despite the support they have received, the literacy skills lagged substantially behind the skills of their peers.

Our findings give weak support to the continuity of familial risk. The means of the NoDys_FR group fell between those of the Dys_FR group and of the control group, but the NoDys-FR and

control groups differed significantly in only one of the reading and spelling measures: text reading accuracy in Grade 3. Note also that the NoDys-FR group performed consistently better than the Dys_FR group. However, although the difference between the NoDys_FR group and the control group was not significant overall, the moderate effect sizes suggest that with a larger sample size, we might have found significant difference. On the other hand, significant differences between groups with or without familial risk and no dyslexia have been found with much smaller sample sizes in English (Snowling et al., 2003, 2007) and in Dutch (Boets et al., 2010; van Bergen et al., 2011).

The majority of the findings of group differences in literacy skills before and at school age in English and Dutch have supported the idea of continuity of familial risk (Boets et al., 2010; Pennington & Lefly, 2001; Snowling et al., 2003, 2007, 2008; van Bergen et al., 2011, 2012), although signs of diminishing group differences along with age have been reported (Boets et al., 2010; Snowling et al., 2007; Torppa et al., 2010; van Bergen et al., 2011). Because in our study, the NoDys_FR group and the control group differed from each other only in one task, no firm conclusions of diminishing versus expanding group differences could be made.

Most prominent support for the continuous nature of familial risk comes from studies employing tasks relying heavily on accurate grapheme-to-phoneme decoding, that is, pseudoword or nonword word reading accuracy (Boets et al., 2010; Pennington & Lefly, 2001; Snowling et al., 2003, 2007; van Bergen et al., 2011). No differences between the two groups with or without familial risk and no dyslexia, on the other hand, have been reported in tasks where other than phonological processing could be used instead or as support of phonological decoding, that is, in word reading (Boets et al., 2010; Snowling et al., 2003; van Bergen et al., 2011) and reading comprehension in adolescence (Snowling et al., 2007). No differences have been reported either in reading task, where there has been no time pressure (i.e., untimed nonword reading accuracy; Snowling et al., 2007) or where the orthography of the language used has been extremely transparent, as in Finnish (Torppa et al., 2010). Therefore, it seems reasonable to think that the requirements of the task or the transparency of orthography or both might affect the visibility of the continuity of familial risk. Finnish is in the shallowest end of the orthographic depth continuum, with close one-to-one correspondence between graphemes and phonemes. This high correspondence makes the learning of decoding and foundation level reading easy (Seymour, Aro, & Erskine, 2003), and most of the children, even those with familial risk, can learn accurate decoding by the end of Grade 2. In English, on the other hand, where nonword reading has been found to be poorer than in more transparent German (Frith et al., 1998), the complexity and inconsistencies of orthography could bring out differences between groups.

The discrepant results concerning the continuous nature of familial risk can also be a consequence of differences in classifying children with or without dyslexia. Whereas in our study, we based the classification on reading speed, reading accuracy, and spelling, Snowling et al. (2003, 2007) based their classification on a composite score that included reading comprehension in addition to word reading and spelling accuracy. It is thus possible that slow readers with good comprehension skills, a group shown to be present at least in the Finnish sample (Torppa et al., 2007), might have ended up in the nondyslexia group. Likewise, van Bergen et

al. (2011) based their classification of children solely on fluency. That is, they did not take reading or spelling inaccuracy as criteria. So, it is also possible that the group of at-risk nondyslexic children in the Dutch sample included children with difficulties in accuracy but not in fluency. This possibility is supported by the finding that in that study, children with typical reading skills but with the familial risk differed from control children only in pseudoword reading in Grade 5 (van Bergen et al., 2011), a task that relies heavily on accurate grapheme-phoneme decoding ability. Interestingly, in another Dutch-speaking sample, the family-risk nondyslexia and control children were more similar to each other when the classification was based on word reading fluency, word reading accuracy, and spelling accuracy (Boets et al., 2010). To further explore this question, researchers should re-analyze the existing data sets applying uniform criteria in classification of children into subgroups.

To better understand the slow reading speed in our Dys_FR group, we compared reading speed in different tasks across groups. In Grades 2 and 3, children in the Dys_FR group read pseudoword texts and word lists at equal speeds, whereas the two other groups read word lists about 1–2 letters/second faster than pseudoword texts. This raises at least two potential suggestions for conclusions. First, it might suggest that children in the Dys_FR group used the same processes in word and pseudoword reading, relying mainly on letter-by-letter decoding. This conclusion is in line with the findings of Ziegler et al. (2003) regarding English- and German-speaking children with dyslexia. In orthographically transparent Italian, Zoccolotti, et al. (2005) have found that Italian children with dyslexia showed a clear word-length effect in word reading, which suggests that the children were still using a sublexical reading procedures in Grade 3. However, Barca, Burani, Di Filippo, and Zoccolotti (2006) reported in an Italian sample that by Grade 6 lexical reading appeared to be available even for children with dyslexia. In the sample of our study, a similar addition of lexical reading process seems to have taken place by Grade 8: children in the Dys_FR group read word lists about 2.5 letters/second faster than pseudoword texts, similar to findings for the children in the two other groups, albeit with the overall lower speed. Second, the slower reading speed of the Dys_FR group in the word-list reading compared with the other two groups can be a consequence not only of poor decoding skills but also of difficulties in the use of orthographic lexicon, as suggested by Bergmann and Wimmer (2008). These difficulties could result from their lower level of exposure to printed text and as a consequence less familiarity with the presented words; word frequency has been shown to have a strong effect on word recognition speed already in school-age children (Zoccolotti et al., 2009). Children in the NoDys_FR and Control groups seemed able to take advantage of their orthographic lexicon and recognize at least the most frequent and therefore familiar words by sight already in Grade 2. This is in line with the findings in another orthographically transparent language, Italian, where lexicality effect was already present in children for high-frequency words at the end of Grade 1 and for low frequency words in Grade 3 (Zoccolotti et al., 2009).

In Grade 2, a similar kind of developmental lag seemed to be present also in the ability of children in the Dys_FR group to use contextual cues, such as syntactic and semantic information: the difference in reading speed between word lists and texts was smaller in the Dys_FR group than in the NoDys_FR and control

groups. At the end of Grade 2, decoding is still difficult in the Dys_FR group, and therefore fewer cognitive resources are left for language processing (Kim et al., 2012; Perfetti, 1985) in children with dyslexia. In Grades 3 and 8, all groups read texts approximately 2 letters/second faster than word list, a result that is in line with the earlier findings in which the same words in context were read faster than without context by fourth graders (Jenkins et al., 2003). Children with dyslexia were beginning to utilize contextual cues from Grade 3, at least 1 year later than normally developing children. And finally, the smaller difference throughout the follow-up period in reading speed between text and pseudoword text reading in the Dys_FR group suggests long-standing deficiencies in automatization of decoding in familiar words, as suggested by Share (2008), or deficient use of word and subword level representations and contextual cues (Snowling, 2008; Stanovich, 1980), or both. Methodological limitations, such as more than one varying factor in comparisons between the tasks, prevent us from making firm conclusions about the processes used in different tasks and to what extent they are specifically compromised in the Dys_FR group.

In conclusion, the findings of the current longitudinal study confirm that the literacy difficulties of children with familial risk for dyslexia and dyslexia in Grade 2 are often persistent. On the other hand, in spite of the familial risk, children who have acquired the basic reading skills follow, for the most part, the developmental track of children without reading difficulties or familial risk later on. In other words, it appears, at least on the group level, that if there are no signs of reading difficulties in Grade 2, one can anticipate typical literacy development also in later grades. But is this true also at the individual level? Do the age-appropriate literacy skills shown here guarantee that these children with familial risk of dyslexia also have age-appropriate reading comprehension skills later, as shown by Snowling et al. (2007) with English-speaking children? These remain important questions for future studies.

References

Abbott, R. D., Berninger, V. W., & Fayol, M. (2010). Longitudinal relationships of levels of language in writing and between writing and reading in Grades 1 to 7. *Journal of Educational Psychology, 102*, 281–298. doi:10.1037/a0019318

Aro, M., & Wimmer, H. (2003). Learning to read: English in comparison to six more regular orthographies. *Applied Psycholinguistics, 24*, 621–635. doi:10.1017/S0142716403000316

Aunola, K., Leskinen, E., Onatsu-Arvilommi, T., & Nurmi, J.-E. (2002). Three methods for studying developmental change: A case of reading skills and self-concept. *British Journal of Educational Psychology, 72*, 343–364. doi:10.1348/000709902320634447

Barca, L., Burani, C., Di Filippo, G., & Zoccolotti, P. (2006). Italian developmental dyslexic and proficient readers: Where are the differences? *Brain and Language, 98*, 347–351. doi:10.1016/j.bandl.2006.05.001

Barker, T. A., Torgesen, J. K., & Wagner, R. K. (1992). The role of orthographic processing skills on five different reading tasks. *Reading Research Quarterly, 27*, 334–345. doi:10.2307/747673

Bergmann, J., & Wimmer, H. (2008). A dual-route perspective on poor reading in a regular orthography: Evidence from phonological and orthographic lexical decisions. *Cognitive Neuropsychology, 25*, 653–676. doi:10.1080/02643290802221404

Bishop, D. V. M. (2009). Genes, cognition, and communication insights from neurodevelopmental disorders. *Annals of New York Academy of Sciences, 1156*, 1–18. doi:10.1111/j.1749-6632.2009.04419.x

Boets, B., De Smedt, B., Cleuren, L., Vandewalle, E., Wouters, J., & Ghesquière, P. (2010). Towards a further characterization of phonological and literacy problems in Dutch-speaking children with dyslexia. *British Journal of Developmental Psychology, 28*, 5–31. doi:10.1348/026151010X485223

Caravolas, M., Hulme, C., & Snowling, M. J. (2001). The foundations of spelling ability: Evidence from a 3-year longitudinal study. *Journal of Memory and Language, 45*, 751–774. doi:10.1006/jmla.2000.2785

Coltheart, M., Rastle, K., Perry, C., Langdon, R., & Ziegler, J. (2001). DRC: A dual route cascaded model of visual word recognition and reading aloud. *Psychological Review, 108*, 204–256. doi:10.1037/0033-295X.108.1.204

Constantinidou, M., & Stainthorp, R. (2009). Phonological awareness and reading speed deficits in reading disabled Greek-speaking children. *Educational Psychology, 29*, 171–186. doi:10.1080/01443410802613483

de Jong, P. F., & van der Leij, A. (2002). Effects of phonological abilities and linguistic comprehension on the development of reading. *Scientific Studies of Reading, 6*, 51–77. doi:10.1207/S1532799XSSR0601_03

de Jong, P. F., & van der Leij, A. (2003). Developmental changes in the manifestation of a phonological deficit in dyslexic children learning to read a regular orthography. *Journal of Educational Psychology, 95*, 22–40. doi:10.1037/0022-0663.95.1.22

Ehri, L. (2005). Learning to read words: Theory, findings, and issues. *Scientific Studies of Reading, 9*, 167–188. doi:10.1207/s1532799xssr0902_4

Fayol, M., Zorman, M., & Létet, B. (2009). Associations and dissociations in reading and spelling French: Unexpectedly poor and good spellers. *British Journal of Educational Psychology Monograph Series II: Pedagogy—Teaching for Learning, 6*, 63–75. doi:10.1348/000709909X421973

Francis, D. J., Shaywitz, S. E., Stuebing, K. K., Shaywitz, B. A., & Fletcher, J. M. (1996). Developmental lag versus deficit models of reading disability: A longitudinal, individual growth curves analysis. *Journal of Educational Psychology, 88*, 3–17. doi:10.1037/0022-0663.88.1.3

Frith, U., Wimmer, H., & Landerl, K. (1998). Differences in phonological recoding in German- and English-speaking children. *Scientific Studies of Reading, 2*, 31–54. doi:10.1207/s1532799xssr0201_2

Furnes, B., & Samuelsson, S. (2010). Predicting reading and spelling difficulties in transparent and opaque orthographies: A comparison between Scandinavian and US/Australian children. *Dyslexia, 16*, 119–142. doi:10.1002/dys.401

Galaburda, A. M., LoTurco, J., Ramus, F., Fitch, R. H., & Rosen, G. D. (2006). From genes to behavior in developmental dyslexia. *Nature Neuroscience, 9*, 1213–1217. doi:10.1038/nn1772

Gallagher, A., Frith, U., & Snowling, M. J. (2000). Precursors of literacy delay among children at genetic risk of dyslexia. *Journal of Child Psychology and Psychiatry, 41*, 203–213. doi:10.1017/S0021963099005284

Giraud, A.-L., & Ramus, F. (2012). Neurogenetics and auditory processing in developmental dyslexia. *Current Opinion in Neurobiology, 23*, 37–42. doi:10.1016/j.conb.2012.09.003

Häyrynen, T., Serenius-Sirve, S., & Korkman, M. (1999). Lukilasse. Lukemisen, kirjoittamisen ja laskemisen seulontatesti ala-asteen luokille 1–6 [Screening test for reading, spelling and counting for the Grades 1–6]. Helsinki, Finland: Psykologien Kustannus Oy.

Hyönä, J. (2011). Foveal and parafoveal processing during reading. In S. Liversedge, I. Gilchrist, & S. Everling (Eds.), *Oxford handbook of eye movements* (pp. 819–838). Oxford, England: Oxford University Press.

Ise, E., & Schulte-Körne, G. (2010). Spelling deficits in dyslexia: Evaluation of an orthographic spelling training. *Annals of Dyslexia*, 60, 18–39. doi:10.1007/s11881-010-0035-8

Jenkins, J. R., Fuchs, L. S., van den Broek, P., Espin, C., & Deno, S. L. (2003). Accuracy and fluency in list and context reading of skilled and RD groups: Absolute and relative performance levels. *Learning Disabilities Research & Practice*, 18, 237–245. doi:10.1111/1540-5826.00078

Kim, Y.-S., & Petscher, Y. (2011). Relations of emergent literacy skill development with conventional literacy skill development in Korean. *Reading and Writing*, 24, 635–656. doi:10.1007/s11145-010-9240-4

Kim, Y.-S., Wagner, R. K., & Lopez, D. (2012). Developmental relations between reading fluency and reading comprehension: A longitudinal study from Grade 1 to Grade 2. *Journal of Experimental Child Psychology*, 113, 93–111. doi:10.1016/j.jecp.2012.03.002

Kuhn, M. R., Schwanenflugel, P. J., & Meisinger, E. B. (2010). Aligning theory and assessment of reading fluency: Automaticity, prosody, and definitions of fluency. *Reading Research Quarterly*, 45, 230–251. doi:10.1598/RRQ.45.2.4

Landerl, K., & Wimmer, H. (2008). Development of word reading fluency and spelling in a consistent orthography: An 8-year follow-up. *Journal of Educational Psychology*, 100, 150–161. doi:10.1037/0022-0663.100.1.150

Landerl, K., Wimmer, H., & Frith, U. (1997). The impact of orthographic consistency on dyslexia: A German–English comparison. *Cognition*, 63, 315–334. doi:10.1016/S0010-0277(97)00005-X

Leinonen, S., Müller, K., Leppänen, P. H. T., Aro, M., Ahonen, T., & Lyytinen, H. (2001). Heterogeneity in adult dyslexic readers: Relating processing skills to the speed and accuracy of oral text reading. *Reading and Writing*, 14, 265–296. doi:10.1023/A:1011117620895

Leppänen, P. H. T., Hämäläinen, J. A., Salminen, H. K., Eklund, K. M., Guttorm, T. K., Lohvansuu, K., . . . Lyytinen, H. (2010). Newborn brain event-related potentials revealing atypical processing of sound frequency and the subsequent association with later literacy skills in children with familial dyslexia. *Cortex*, 46, 1362–1376. doi:10.1016/j.cortex.2010.06.003

Leppänen, U., Niemi, P., Aunola, K., & Nurmi, J.-E. (2004). Development of reading skills among preschool and primary school pupils. *Reading Research Quarterly*, 39, 72–93. doi:10.1598/RRQ.39.1.5

Leppänen, U., Niemi, P., Aunola, K., & Nurmi, J.-E. (2006). Development of reading and spelling Finnish from preschool to Grade 1 and Grade 2. *Scientific Studies of Reading*, 10, 3–30. doi:10.1207/s15327999ssr1001_2

Lervåg, A., & Hulme, C. (2010). Predicting the growth of early spelling skills: Are there heterogeneous developmental trajectories? *Scientific Studies of Reading*, 14, 485–513. doi:10.1080/10888431003623488

Lyon, G. R., Shaywitz, S. E., & Shaywitz, B. A. (2003). A definition of dyslexia. *Annals of Dyslexia*, 53, 1–14. doi:10.1007/s11881-003-0001-9

Lyytinen, H., Erskine, J., Ahonen, T., Aro, M., Eklund, K., Guttorm, T., . . . Virolainen, H. (2008). Early identification and prevention of dyslexia: Results from a prospective follow-up study of children at familial risk for dyslexia. In G. Reid, F. Manis, & L. Siegel (Eds.), *The Sage handbook of dyslexia* (pp. 121–146). Thousand Oaks, CA: Sage.

Moll, K., & Landerl, K. (2009). Double dissociation between reading and spelling deficits. *Scientific Studies of Reading*, 13, 359–382. doi:10.1080/10888430903162878

Parrila, R., Aunola, K., Leskinen, E., Nurmi, J.-E., & Kirby, J. R. (2005). Development of individual differences in reading: Results from longitudinal studies in English and Finnish. *Journal of Educational Psychology*, 97, 299–319. doi:10.1037/0022-0663.97.3.299

Pennala, R., Eklund, K., Hämäläinen, J., Richardson, U., Martin, M., Leivo, M., . . . Lyytinen, H. (2010). Perception of phonemic length and its relation to reading and spelling skills in children with familial risk for dyslexia at the three first grades in school. *Journal of Speech, Language, and Hearing Research*, 53, 710–724. doi:10.1044/1092-4388(2009/08-0133)

Pennington, B. F. (2006). From single to multiple deficit models of developmental disorders. *Cognition*, 101, 385–413. doi:10.1016/j.cognition.2006.04.008

Pennington, B. F., & Lefly, D. L. (2001). Early reading development in children at family risk for dyslexia. *Child Development*, 72, 816–833. doi:10.1111/1467-8624.00317

Pennington, B. F., Santerre-Lemmon, L., Rosenberg, J., MacDonald, B., Boada, R., Friend, A., . . . Olson, R. K. (2012). Individual prediction of dyslexia by single versus multiple deficit models. *Journal of Abnormal Psychology*, 121, 212–224. doi:10.1037/a0025823

Perfetti, C. A. (1985). *Reading ability*. New York, NY: Oxford University Press.

Posner, M. I., & Snyder, C. R. R. (1975). Attention and cognitive control. In R. Solso (Ed.), *Information processing and cognition: The Loyola Symposium* (pp. 55–85). Hillsdale, NJ: Erlbaum.

Puolakanaho, A., Ahonen, T., Aro, M., Eklund, K., Leppänen, P. H. T., Poikkeus, A.-M., . . . Lyytinen, H. (2007). Very early phonological and language skills: Estimating individual risk of reading disability. *Journal of Child Psychology and Psychiatry*, 48, 923–931. doi:10.1111/j.1469-7610.2007.01763.x

Puolakanaho, A., Ahonen, T., Aro, M., Eklund, K., Leppänen, P. H. T., Poikkeus, A.-M., . . . Lyytinen, H. (2008). Developmental links of very early phonological and language skills to the second grade reading outcomes: Strong to accuracy but only minor to fluency. *Journal of Learning Disabilities*, 41, 353–370. doi:10.1177/0022219407311747

Raven, J. C., Court, J. H., & Raven, J. (1992). *Standard progressive matrices*. Oxford, England: Oxford Psychologists Press.

Scarborough, H. S. (1990). Very early language deficits in dyslexic children. *Child Development*, 61, 1728–1743. doi:10.2307/1130834

Scerri, T. S., & Schulte-Körne, G. (2010). Genetics of developmental dyslexia. *European Child & Adolescent Psychiatry*, 19, 179–197. doi:10.1007/s00787-009-0081-0

Seymour, P. H. K. (1995). *Cognitive Workshop, Version 1.1: User manual*. Dundee, United Kingdom: University of Dundee.

Seymour, P. H. K., Aro, M., & Erskine, J. M. (2003). Foundation literacy acquisition in European orthographies. *British Journal of Psychology*, 94, 143–174. doi:10.1348/000712603321661859

Share, D. L. (2008). On the Anglocentricities of current reading research and practice: The perils of overreliance on an “outlier” orthography. *Psychological Bulletin*, 134, 584–615. doi:10.1037/0033-2909.134.4.584

Shaywitz, S. E., Fletcher, J. M., Holahan, J. M., Schneider, A. E., Marichione, K. E., Stuebing, K. K., . . . Shaywitz, B. A. (1999). Persistence of dyslexia: The Connecticut Longitudinal Study at Adolescence. *Pediatrics*, 104, 1351–1359.

Shaywitz, B. A., Holford, T. R., Holahan, J. M., Stuebing, K. K., Francis, D. J., & Shaywitz, S. E. (1995). A Matthew effect for IQ but not for reading: Results from a longitudinal study. *Reading Research Quarterly*, 30, 894–906. doi:10.2307/748203

Snowling, M. J. (2008). Specific disorders and broader phenotypes: The case of dyslexia. *Quarterly Journal of Experimental Psychology*, 61, 142–156. doi:10.1080/17470210701508830

Snowling, M. J., Adams, J. W., Bishop, D. V. M., & Stothard, S. E. (2001). Educational attainments of school leavers with a preschool history of speech-language impairments. *International Journal of Language and Communication Disorders*, 36, 173–183. doi:10.1080/13682820010019892

Snowling, M. J., Callagher, A., & Frith, U. (2003). Family risk of dyslexia is continuous: Individual differences in the precursors of reading skill. *Child Development*, 74, 358–373. doi:10.1111/1467-8624.7402003

Snowling, M. J., Muter, V., & Carroll, J. (2007). Children at family risk of dyslexia: A follow-up in early adolescence. *Journal of Child Psychology and Psychiatry*, 48, 609–618. doi:10.1111/j.1469-7610.2006.01725.x

Stanovich, K. E. (1980). Toward an interactive-compensatory model of individual differences in the development of reading fluency. *Reading Research Quarterly*, 16, 32–71. doi:10.2307/747348

Torppa, M., Lyytinen, P., Erskine, J., Eklund, K., & Lyytinen, H. (2010). Language development, literacy skills, and predictive connections to reading in Finnish children with and without familial risk for dyslexia. *Journal of Learning Disabilities*, 43, 308–321. doi:10.1177/0022219410369096

Torppa, M., Parrila, R., Niemi, P., Lerkkanen, M.-K., Poikkeus, A. M., & Nurmi, J.-E. (2013). The double deficit hypothesis in the transparent Finnish orthography: A longitudinal study from kindergarten to Grade 2. *Reading and Writing*, 26, 1353–1380. doi:10.1007/s11145-012-9423-2

Torppa, M., Tolvanen, A., Poikkeus, A.-M., Eklund, K., Lerkkanen, M.-K., Leskinen, E., & Lyytinen, H. (2007). Reading development subtypes and their early characteristics. *Annals of Dyslexia*, 57, 3–32. doi:10.1007/s11881-007-0003-0

Vaessen, A., & Blomert, L. (2010). Long-term cognitive dynamics of fluent reading development. *Journal of Experimental Child Psychology*, 105, 213–231. doi:10.1016/j.jecp.2009.11.005

van Bergen, E., de Jong, P. F., Plakas, A., Maassen, B., & van der Leij, A. (2012). Child and parental literacy levels within families with a history of dyslexia. *Journal of Child Psychology and Psychiatry*, 53, 28–36. doi:10.1111/j.1469-7610.2011.02418.x

van Bergen, E., de Jong, P. F., Regtvoort, A., Oort, F., van Otterloo, S., & van der Leij, A. (2011). Dutch children at family risk of dyslexia: Precursors, reading development, and parental effects. *Dyslexia*, 17, 2–18. doi:10.1002/dys.423

Wechsler, D. (1991). *Wechsler Intelligence Scales for Children* (3rd ed.). Sidcup, England: The Psychological Corporation.

Wimmer, H. (1996). The nonword reading deficit in developmental dyslexia: Evidence from children learning to read German. *Journal of Experimental Child Psychology*, 61, 80–90. doi:10.1006/jecp.1996.0004

Wimmer, H., & Mayringer, H. (2002). Dysfluent reading in the absence of spelling difficulties: A specific disability in regular orthographies. *Journal of Educational Psychology*, 94, 272–277. doi:10.1037/0022-0663.94.2.272

Ziegler, J. C., Perry, C., Ma-Wyatt, A., Ladner, D., & Schulte-Körne, G. (2003). Developmental dyslexia in different languages: Language-specific or universal? *Journal of Experimental Child Psychology*, 86, 169–193. doi:10.1016/S0022-0965(03)00139-5

Zoccolotti, P., De Luca, M., Di Filippo, G., Judica, A., & Martelli, M. (2009). Reading development in an orthographically regular language: Effects of length, frequency, lexicality, and global processing ability. *Reading and Writing*, 22, 1053–1079. doi:10.1007/s11145-008-9144-8

Zoccolotti, P., De Luca, M., Di Pace, E., Gasperini, F., Judica, A., & Spinelli, D. (2005). Word length effect in early reading and in developmental dyslexia. *Brain and Language*, 93, 369–373. doi:10.1016/j.bandl.2004.10.010

Zoccolotti, P., De Luca, M., Di Pace, E., Judica, A., Orlandi, M., & Spinelli, D. (1999). Markers of developmental surface dyslexia in a language (Italian) with high grapheme-phoneme correspondence. *Applied Psycholinguistics*, 20, 191–216. doi:10.1017/S0142716499002027

Received January 16, 2013

Revision received April 4, 2014

Accepted April 23, 2014 ■

Parallel and Serial Reading Processes in Children's Word and Nonword Reading

Madelon van den Boer and Peter F. de Jong
University of Amsterdam

Fluent reading is characterized by rapid and accurate identification of words. It is commonly accepted that such identification relies on the availability of orthographic knowledge. However, whether this orthographic knowledge should be seen as an accumulation of word-specific knowledge in a lexicon acquired through decoding or as a well-developed associative network of sublexical units is still under debate. We studied this key issue in reading research by looking at the serial and/or parallel reading processes underlying word and nonword reading. Participants were 314 Dutch 2nd, 3rd, and 5th graders. The children were administered digit, word, and nonword naming tasks. We used latent class analyses to distinguish between readers who processed the letter strings serially or in parallel, based on the correlation patterns of word and nonword reading with serial and discrete digit naming. The 2 classes of readers were distinguished for both word and nonword reading. The validity of these classes was supported by differences in sensitivity to word and nonword length. Interestingly, the different classes seemed to reflect a developmental shift from reading all letter strings serially toward parallel processing of words, and later of nonwords. The results are not fully in line with current theories on the representation of orthographic knowledge. Implications in terms of models of the reading process are discussed.

Keywords: decoding, sight word reading, parallel processing, reading development, nonwords

Fluent reading is characterized by rapid and accurate identification of words. Such identification is commonly believed to depend on the availability of orthographic knowledge (e.g., Ehri, 2005; Share, 2008). However, the proper representation of orthographic knowledge in a model of reading is still under debate. On the one hand, it has been proposed that readers acquire word-specific knowledge and store this knowledge in a lexicon (e.g., Coltheart, Rastle, Perry, Langdon, & Ziegler, 2001; Jackson & Coltheart, 2001). Upon encountering familiar words in written form, pronunciation and meaning can immediately and automatically be retrieved from memory (Ehri, 2005). On the other hand, it has been proposed that the reading system is an associative network of interconnected sublexical units, without lexical memory for words (e.g., Plaut, McClelland, Seidenberg, & Patterson, 1996; Seidenberg & McClelland, 1989). First, we discuss these two approaches and their implications in more detail. Next, we consider methods to determine whether word identification is based on the retrieval of pronunciations from memory.

According to the first or word-specific approach, fluent reading means reading by sight. For a word to be read by sight, a connec-

tion must be made between the orthographic form of a word and its previously acquired phonological counterpart (Ehri, 2005). According to the self-teaching hypothesis (Share, 1995, 1999), a reader can acquire the detailed orthographic representations necessary for fast and efficient reading through phonological recoding of novel letter strings. Every time a reader successfully decodes a printed word into a phonological code, an orthographic representation of that word is built or strengthened. Therefore, beginning readers initially rely on decoding skills to read words, but read more fluently when previous encounters with words have accumulated in well-established orthographic representations.

This development of the reading system, from heavy reliance on decoding toward reading an increasing number of words by sight, fits well with the dual route cascaded (DRC) model of reading (Coltheart et al., 2001; Jackson & Coltheart, 2001). Therefore, the DRC model provides a useful framework in studying reading development, although it should be noted that the model is intended to model reading aloud of monosyllabic letter strings by adult fluent readers. Within the DRC model, two routes are distinguished that are simultaneously active. Initial parallel identification of letter identities is common to both routes. Subsequently, phonology is activated through the lexical and nonlexical routes. Sight word reading is represented as reading through the lexical route. In the lexical route, word identification is achieved in parallel by successive activation of the word's entry in the orthographic and phonological lexicon. Decoding, dominating the processing of unfamiliar words or nonwords, is modeled with the nonlexical route. This route works in parallel to the lexical route, but graphemes are serially decoded into phonemes according to grapheme-phoneme conversion rules. As a result of reading experience, one could expect a gradual shift in dominance from the

This article was published Online First June 16, 2014.

Madelon van den Boer and Peter F. de Jong, Research Institute of Child Development and Education, University of Amsterdam.

We thank Marleen Haentjens-van Meeteren for her involvement in task development and the data collection.

Correspondence concerning this article should be addressed to Madelon van den Boer, University of Amsterdam, Research Institute of Child Development and Education, P.O. Box 94208, 1090 GE Amsterdam, the Netherlands. E-mail: m.vandenboer@uva.nl

nonlexical route, when many words are decoded early in development, toward the lexical route, when an increasing number of words become represented in the orthographic lexicon and can be quickly recognized by sight.

An important characteristic of the DRC model is that words can only be read by sight if a word-specific representation is present in the orthographic lexicon (e.g., Coltheart et al., 2001). In other words, reading development is item specific. Orthographic representations can exist only if words have previously been encountered and decoded successfully. And words can be processed in parallel only if orthographic representations exist that are connected to the representations of the same words in the phonological lexicon.

This idea of word-specific orthographic knowledge, however, stands in sharp contrast to the second approach in modeling the reading system. According to the parallel distributed processing model (PDP; e.g., Plaut et al., 1996), for example, word-specific representations do not exist. Rather, letter strings are read by a reading system based on parallel activation of interconnected orthographic, phonological, and semantic units. The interactions among these units are governed by connection weights that represent the system's knowledge of spelling–sound correspondences in the language input. Within this associative network of sublexical units, there is no orthographic or phonological lexicon for words.

As a result of the different representations of orthographic knowledge as either word-specific or sublexical, the DRC and PDP models of reading also have different definitions of fluent reading. Fluent reading, in the DRC model (e.g., Coltheart et al., 2001), entails reading by sight, which occurs through parallel activation of phonology of a letter string by accessing representations in the orthographic and phonological lexicon. In contrast, fluent reading in the PDP model (e.g., Plaut et al., 1996) entails parallel activation of phonology from print via sublexical units. Both models, however, predict that fluent word reading is achieved through parallel computation of phonology from the letter string.

The models differ greatly in how they account for the reading of nonwords. According to the DRC model (e.g., Coltheart et al., 2001), nonwords cannot be represented in the orthographic lexicon, and as a result always require involvement of the nonlexical route. In contrast, PDP models do not presume a separate mechanism for the reading of unfamiliar words and nonwords. According to the PDP model (e.g., Plaut et al., 1996), all letter strings are read by the same reading system through parallel activation of the interconnected units. Nonwords, especially those that adhere to regular orthographic and phonological patterns, are not processed differently from words.

A key issue in distinguishing between these two models of reading is whether phonological codes of words and nonwords are activated in parallel. Within the DRC framework, length effects have been studied as indicators of whether phonology is activated predominantly serially or in parallel. In the early stages of reading development, the speed of single word and nonword reading increases as a function of the number of letters, whereas in advanced readers this length effect becomes restricted to longer words (i.e., more than six letters) and nonwords (e.g., Marinus & de Jong, 2010; Spinelli et al., 2005; van den Boer, de Jong, & Haentjens-van Meeteren, 2013; Weekes, 1997; Ziegler, Perry, Ma-Wyatt, Ladner, & Schulte-Körne, 2003; Zoccolotti et al., 2005). A length effect is presumed to occur when words are

identified through serial activation of phonology, whereas the absence of a length effect indicates that phonology is activated in parallel. In line with the DRC model, length effects remain for nonwords, which are supposed to be read predominantly through the nonlexical route.

However, although a length effect is indeed expected when letter strings are decoded, the reverse—that an observed length effect is the result of decoding—is not necessarily true. In fact, length effects have been found that could not be ascribed to serial processing through the nonlexical route (Risko, Lanthier, & Besner, 2011; van den Boer, de Jong, & Haentjens-van Meeteren, 2012). Risko et al. (2011), for example, found that increased spacing between letters resulted in increased effects of item length. These effects, however, were found at the level of letter identification, not serial activation of phonology. Similarly, Van den Boer et al. (2012) found length effects in the lexical decisions of children, while independent evidence suggested that items were processed in parallel through the lexical route. Together, these findings indicate that a length effect in and of itself does not prove that serial processes underlie word identification. Moreover, in PDP models, length effects are not interpreted in terms of decoding but are ascribed to other factors, such as visual and articulatory factors or differences in orthographic neighborhood size (Seidenberg & Plaut, 1998; but see Plaut, 1999, for an attempt to model length effects within a PDP framework). Thus, length effects are expected when words are decoded, but a length effect in itself does not prove that words have been identified through serial decoding. Additional independent evidence for a serial or parallel reading strategy is called for.

As an alternative, it has been proposed that parallel processing can be detected by the speed with which single words are read. Ehri and Wilce (1983) compared how quickly beginning readers could identify highly familiar, overlearned symbols (i.e., digits) with the readers' word recognition speed. In skilled readers, response latencies to both digits and words were equal as early as first grade. In less skilled readers, however, similar response rates were obtained later, around third or fourth grade. These results indicated that even in the first years of reading development, words are no longer decoded but are processed in parallel and elicit the same routinized naming responses as overlearned symbols. Interestingly, Ehri and Wilce (1983) also included three-letter nonwords in their study and found that skilled readers also identified these nonwords as quickly as digits. Less skilled readers, however, identified nonwords slower than digits at least up to fourth grade. These findings suggest that nonword phonology could potentially be activated in parallel.

In line with Ehri and Wilce (1983); Aaron et al. (1999) showed that if a word is processed in parallel, the speed of reading this word is close to the speed of naming letters. Similar results have also been reported by van den Bos, Zijlstra, and Van den Broeck (2003), who showed that naming speed of alphanumeric symbols (i.e., letters and digits) was closely related to monosyllabic word naming speed. However, naming speed is greatly influenced by word frequency (e.g., Forster & Chambers, 1973; Frederiksen & Kroll, 1976). The phonological codes of digits are very frequent, which results in relatively short naming latencies. Therefore, similar reading latencies to digits and words are probably only found when high frequency words are studied. Reading latencies to

words of lower frequency might not be equal to reading latencies of digits, even though these words might be processed in parallel.

To get around this problem, de Jong (2011) argued that if word reading relies on a parallel retrieval process, individual differences in word reading and digit naming speed should be similar. Therefore, a high correlation should be found between word and digit naming, despite possible differences in absolute naming speed. More specifically, de Jong proposed to consider the relations of serial and discrete digit naming with word reading to determine whether a particular set of words is read by sight. Digit naming concerns the rapid naming of digits. Whereas in serial naming the digits are presented in rows, in discrete naming digits are presented one by one, on a computer screen. Naming latencies of digits presented in a discrete format were assumed to reflect lexical access speed, the retrieval of known phonological codes from memory. If words, also presented in a discrete format, are processed in parallel, a high correlation is expected with discrete digit naming. If, however, words are read through decoding, a stronger relation could be expected with a serial format of digit naming, because both activation of phonology and serial digit naming reflect a serial process. The correlation patterns were in line with both of these expectations in showing that for beginning readers (Grade 1), who are expected to rely predominantly on decoding, word reading was most strongly related to serial digit naming, whereas discrete digit naming was the stronger correlate for more advanced readers, who are expected to process short words in parallel (Grades 2 and 4).

As a next step, de Jong (2011) showed through latent class analyses that the children from the three grades could be assigned to two classes of readers. For a large class of readers, single word reading related strongly to discrete digit naming. For a second, smaller class of readers, however, single word reading related more strongly to serial digit naming. This suggested that the first class of readers processed the words in parallel, similar to naming a digit. The second class of readers, however, was not processing the words in parallel but predominantly relied on a serial decoding strategy instead. De Jong argued that this classification is fully compatible with an item-specific view of reading development, such as the DRC model (e.g., Coltheart et al., 2001). Whether a reader processed the words in parallel or serially depended on whether the words in the set were represented in the lexicon or not. If the words were represented in the lexicon, the words were read by sight. If the majority of the words in the set were not represented in the lexicon, the main reading strategy would be serial decoding. In other words, the classifications depend on the words that were presented. The number of classes would vary with the number of word sets used, and the sizes of the classes with the difficulty of the words included.

In the current study we focused on word and nonword reading in Grades 2, 3, and 5. For word reading, we expected to find two classes of readers, namely, serial and parallel processors. More importantly, we studied whether these results are tied to a particular set of words by studying whether similar classes of readers could be distinguished for nonword reading. According to an item-specific view of reading development, and in line with the DRC model, all readers should have a predominantly serial reading strategy for nonwords; thus, only one class of serial nonword readers should be identified. These predictions are tested against the predictions of the PDP model (e.g., Plaut et al., 1996), which

states that both words and nonwords are processed in parallel by all readers. Thus, a single class of parallel processors would be expected for both word and nonword reading. If nonwords, like words, can be processed in parallel, this would indicate that serial and parallel reading processes were not tied to particular sets of words but could potentially be generalized to all short words and nonwords. A second novel aspect in the current study is the focus on validating the interpretation of the different classes of readers by examining length effects. Reading latencies of serial processors are expected to be affected by word length, whereas the reading latencies of parallel processors are hypothesized to be independent of length.

Method

Participants

A total of 314 Dutch children participated in the study. One hundred seventeen children attended second grade (52 boys, 65 girls), 86 third grade (44 boys, 42 girls), and 111 fifth grade (51 boys, 60 girls). The mean ages of the children were 8 years ($SD = 5.70$ months) in Grade 2, 9 years 4 months ($SD = 6.58$ months) in Grade 3, and 11 years ($SD = 5.86$ months) in Grade 5. All children attended mainstream primary education. Scores on the One Minute Reading Test (*Eén Minuut Test*; Brus & Voeten, 1995), a standardized test of word reading fluency with an average of 10 and a standard deviation of 3, showed that the sample included a representative range of reading abilities (Grade 2: $M = 10.66$, $SD = 2.93$; Grade 3: $M = 10.54$, $SD = 2.52$; Grade 5: $M = 9.25$, $SD = 2.58$). All children had normal or corrected to normal vision.

Measures

A word and nonword reading task was administered to all children, as well as serial and discrete measures of digit naming.

Discrete word and nonword reading. The reading task consisted of 45 words and 45 nonwords varying in length from three to five letters. For each length, 15 monosyllabic words were selected from a corpus of child literature of two million tokens (Schrooten & Vermeer, 1994). Across lengths, words were matched on onset (i.e., the first letter) and frequency. The words ranged in frequency to reflect the variation in words children encounter ($Mdn = 23$, range: 1–148). Nonwords were created by interchanging onsets and rhymes of the words. For example, the words *drift*, *front*, and *kramp* (meaning *urge*, *front*, and *cramp*, respectively) were used to create the nonwords *dront*, *framp*, and *krift*. Therefore, words and nonwords were matched on onset and consonant–vowel structure. When the created nonword was unpronounceable or also a Dutch word, one letter was changed in the rhyme.

The reading task (as well as the discrete digit-naming task described below) was programmed in E-prime (Version 1.0; Schneider, Eschman, & Zuccolotto, 2002). Words and nonwords were presented one by one in the middle of a laptop screen (14.1 in.; 35.8 cm) in 72-point Arial font. A plus sign presented for 750 ms focused attention. Then the word or nonword appeared, and children were asked to read it aloud as quickly and accurately as possible. A voice key registered naming latencies from the onset of stimulus presentation until the onset of the response. The experi-

menter registered naming accuracy on a response box (correct and valid, incorrect, or invalid). Words and nonwords were presented in blocks, separated by a fixed break of 1.5 min. The order of word and nonword reading was counterbalanced across the children.

Digit naming. Naming of digits (1, 3, 5, 6, and 8) was administered in serial and discrete format.

Serial digit naming. The five digits were presented 10 times in a random order on a sheet with five lines of 10 digits each (see Denckla & Rudel, 1976). Children were asked to name aloud all digits as quickly as possible. The time needed to name all 50 digits was converted to the number of digits named per second.

Discrete digit naming. The 50 digits were also presented in a discrete naming task, in the same order as in the serial task. The digits were presented one by one in the middle of a laptop screen (14.1 in.; 35.8 cm) in 72-point Arial font. Each trial started with a plus sign, presented for 750 ms, to focus attention. Then the digit was presented and remained on the screen until the child made a response. A voice key registered response latencies from the onset of presentation until the onset of the response. The experimenter registered naming accuracy on a response box (correct and valid, incorrect, or invalid). The score consisted of the mean naming latency per digit, converted to the number of digits named per second.

Procedure

Children in second and fifth grade were tested in January/February, when they had received approximately 1 year 5 months and 4 years 5 months of reading instruction, respectively. Third graders were tested in June/July, after approximately 3 years of reading instruction, meaning that the reading age of these children lay exactly between the reading ages of second and fifth graders. The word and nonword reading task and the digit naming tasks were administered during two waves of more extensive data collection. Second and fifth graders participated in a classroom session of about 1 hr 30 min and two individual sessions of approximately 30 min each. Third graders completed the experimental tasks during one individual session of approximately 40 min.

Results

Clustering Readers Based on Reading Processes

As to be expected in a transparent orthography, mean accuracy across grades was high for both words ($M = 0.95$, $SD = 0.07$) and nonwords ($M = 0.92$, $SD = 0.09$). Reading latencies were ex-

cluded from analysis if the voice key was not validly triggered (5.9%), if latencies were less than 250 ms or more than 6,000 ms (0.9%), and if latencies were more than 3 standard deviations from a participant's mean (1.6%). Similar to de Jong (2011), word and nonword reading latencies were converted into fluency scores reflecting the number of items read correctly per second. First, average word and nonword latencies were calculated for each child and transformed to the number of items read per second to normalize scores. Then, the proportion of words and nonwords read correctly was calculated over valid trials. Finally, word and nonword fluency scores were calculated by multiplying the number of items read per second by the proportion of items correct. For clarity purposes, we use the terms *word reading fluency* and *nonword reading fluency* to refer to these reading scores. However, please note that the measures of reading fluency are based on discrete word and nonword reading tasks.

Scores on word and nonword reading fluency, as well as on serial and discrete digit naming, were normally distributed in each grade separately and in the entire data set. All variables were inspected for univariate outliers (i.e., a score of more than 3 standard deviations above or below the mean), separately for each grade. Two outliers (one in Grade 3 and one in Grade 5) were identified for word reading, one (in Grade 3) for nonword reading, two (one in Grade 3 and one in Grade 5) for serial digit naming, and two (one in Grade 3 and one in Grade 5) for discrete digit naming. These scores were coded as missing and not included in the analyses. None of the children was identified as a multivariate outlier.

Descriptive statistics. The means and standard deviations on word and nonword reading fluency and serial and discrete digit naming for each grade are shown in Table 1. Overall, growth can be seen across grades. Both reading fluency and digit naming speed increased significantly between Grades 2 and 3. Between Grades 3 and 5, only discrete digit naming speed significantly increased. Across all grades, average reading fluency was lower than average digit naming speed.

Correlations between word and nonword reading fluency and serial and discrete digit naming for each grade are shown in Table 2. In Grade 2, word reading fluency correlated equally strongly with serial and discrete digit naming ($Z = 0.701$, $p = .414$). In Grades 3 and 5, however, word reading was more strongly related to discrete than to serial digit naming (Grade 3: $Z = 2.216$, $p < .05$; Grade 5: $Z = 4.036$, $p < .001$). Interestingly, a similar pattern was found in the correlations between nonword reading fluency and digit naming. In Grade 2, nonword reading fluency was related

Table 1

Means (and Standard Deviations) on Word and Nonword Reading Fluency, and Serial and Discrete Digit Naming in Items per Second in Grades 2, 3, and 5

Variable	Grade 2 (<i>N</i> = 117) <i>M</i> (<i>S</i>)	Grade 3 (<i>N</i> = 86) <i>M</i> (<i>S</i>)	Grade 5 (<i>N</i> = 111) <i>M</i> (<i>S</i>)	<i>t</i> statistics 2 vs. 3	<i>t</i> statistics 3 vs. 5
Word reading fluency	1.13 (.43)	1.62 (.25)	1.68 (.27)	10.284**	1.610
Nonword reading fluency	.99 (.43)	1.44 (.29)	1.46 (.34)	8.780**	0.463
Serial digit naming	1.75 (.39)	2.19 (.42)	2.27 (.45)	7.703**	1.241
Discrete digit naming	1.69 (.26)	1.91 (.25)	2.00 (.31)	6.080**	2.027*

* $p < .05$. ** $p < .01$.

Table 2

Correlations of Word and Nonword Reading Fluency With Serial and Discrete Digit Naming in Grades 2, 3, and 5

Digit naming	Words			Nonwords		
	Grade 2 (N = 117)	Grade 3 (N = 86)	Grade 5 (N = 111)	Grade 2 (N = 117)	Grade 3 (N = 86)	Grade 5 (N = 111)
Serial	.532**	.274*	.232*	.564**	.338**	.343**
Discrete	.467**	.503**	.643**	.454**	.444**	.543**

* $p < .05$. ** $p < .01$.

equally strongly to serial and discrete digit naming ($Z = 1.397$, $p = .163$). In contrast to words, equal relations were also found in Grade 3 ($Z = 1.024$, $p = .306$). In Grade 5, however, the difference of the correlation of nonword reading with discrete and serial digit naming approached significance, in favor of discrete digit naming ($Z = 1.936$, $p = .053$).

A series of stepwise regression analyses was conducted to examine whether serial and discrete digit naming were independent predictors of reading fluency. The analyses were conducted for each grade, with word and nonword reading fluency as dependent variables. In the first set of analyses serial digit naming was entered first, and it was determined whether including discrete digit naming resulted in additional explained variance. In the second set, the order of serial and discrete digit naming was reversed. The (additional) variance explained in each step is presented in Table 3. In Grade 2, both serial and discrete digit naming explained unique variance in word reading fluency. In Grades 3 and 5, however, discrete digit naming was the stronger predictor, and serial digit naming did not explain additional variance. For nonword reading fluency, the results were the same, with the exception of a small independent effect of serial digit naming on nonword reading fluency in Grade 5. Interestingly, the results clearly show an increase in the amount of variance in reading fluency explained by discrete digit naming and a decrease in the amount of variance explained by serial digit naming.

Classes of readers. The correlation patterns and regression results indicate that in the early stages of reading development (i.e., Grade 2) serial digit naming is the stronger correlate and predictor of reading fluency, whereas reading becomes more strongly related to discrete digit naming in the higher grades. This might suggest that two classes of readers could be found: one class for whom word reading is related more strongly to serial digit naming, and one class for whom word reading is related more

strongly to discrete digit naming. Alternatively, three classes of readers could be expected, when readers are better classified by grade. Therefore, both two- and three-class models were fitted and compared. Correlation patterns with nonword reading fluency suggest that similar clusters of children could be found based on the relations between nonword reading and digit naming. Therefore, the same models were estimated based on nonword reading, serial digit naming, and discrete digit naming.¹

If a (categorical) variable is measured that can be the source of heterogeneity within a sample, this variable can be used to split participants into groups, and differences can be analyzed through multiple group analyses. If, however, the source of heterogeneity is hypothesized but unobserved, as are reading strategies in the current study, factor mixture modeling can be used to determine classes within a heterogeneous sample (Lubke & Muthén, 2005). Through factor mixture modeling, participants were clustered into unobserved (latent) classes based on mean scores on and correlations between a set of observed variables. Three variables were input for the current analyses: word or nonword reading, serial digit naming, and discrete digit naming.

Models distinguishing between two classes and three classes were fitted using Mplus (Version 5.21; Muthén & Muthén, 2009). Several statistics can be obtained to evaluate model fit and decide on the number of classes. However, Nylund, Asparouhov, and Muthén (2007) showed that the Bayesian information criterion (BIC) and bootstrap likelihood ratio test (BLRT) should be favored. Models with lower BIC values should be preferred. The BLRT p value indicates whether a model with k classes significantly improves fit over a model with $k - 1$ classes. In addition, entropy was used to evaluate the models, with a value close to 1 indicating low average likelihoods that a child assigned to one class could have been assigned to another (Celeux & Soromenho, 1996).

For the word reading fluency models, the two-class model was favored over the three-class model, according to BIC (two classes: 633.19; three classes: 648.42) and entropy (two classes: .878; three classes: .756). In addition, the BLRT indicated that the two-class model fitted significantly better than a one-class model ($p < .001$), but that a three-class model did not significantly improve fit over a two-class model ($p = .92$). The results of the two-class solution are presented in Table 4. For a large class of 277 children, word

Table 3

R² Changes in Hierarchical Regression Analyses Using Serial and Discrete Rapid Naming to Predict Word and Nonword Reading Fluency

Digit naming	Words			Nonwords		
	Grade 2	Grade 3	Grade 5	Grade 2	Grade 3	Grade 5
1. Serial	.28**	.08*	.04*	.32**	.11**	.10**
2. Discrete	.06**	.18**	.39**	.05**	.11**	.24**
1. Discrete	.22**	.26**	.42**	.21**	.20**	.30**
2. Serial	.13**	.00	.01	.16**	.02	.04**

* $p < .05$. ** $p < .01$.

¹ Including word and nonword reading in one mixture model resulted in classes that did not fit expected correlation patterns and were difficult to interpret. This is possibly due to the high correlation between word and nonword reading.

Table 4

Correlations of Serial and Discrete Digit Naming With Word and Nonword Reading Fluency in Classes of Readers

Digit naming	Word reading fluency		Nonword reading fluency	
	Serial processors (N = 37)	Parallel processors (N = 277)	Serial processors (N = 69)	Parallel processors (N = 245)
Serial	.551*	.438*	.545*	.403*
Discrete	.462*	.674*	.483*	.613*

* $p < .01$.

reading correlated more strongly with discrete than with serial digit naming, suggesting that words are processed in parallel or read by sight. Children from each grade were assigned to this class of parallel processors (83 second, 84 third, and 110 fifth graders). However, for a smaller class of 37 children, word reading was most strongly related to serial digit naming, suggesting that words are not (yet) processed in parallel. This class of serial processors consisted mainly of children in Grade 2 (34 second graders, 2 third graders, and 1 fifth grader).

For the nonword reading fluency models, the two-class model was favored over the three-class model according to BIC (two classes: 690.22; three classes: 704.49) but not according to entropy (two classes: .775; three classes: .836). The BLRT, however, indicated that the two-class model fitted significantly better than a one-class model ($p < .001$), but that a three-class model did not significantly improve fit over a two-class model ($p = .89$). Moreover, one of the classes in the three-class solution included only nine children, and the interrelations among the variables within the classes were difficult to interpret. Therefore, the two-class solution seemed best. The results of the two-class solution are presented in Table 4. In line with the result for word reading, nonword reading correlated more strongly with discrete than serial digit naming for a large class of 245 children, suggesting that nonwords were processed in parallel. Parallel processors were identified in each grade (66 second, 79 third, and 100 fifth graders). For a smaller class of 69 children, nonword reading related more strongly to serial digit naming. This class of serial processors, who did not (yet) process nonwords in parallel, consisted mainly of children in Grade 2, although small groups of third and fifth graders were also assigned to this class (51 second graders, 7 third graders, and 11 fifth graders).

Length Effects

If our interpretation of the classes of readers is correct, differences would be expected across classes in length effects. Length effects are expected when letter strings are processed serially. Therefore, length effects were expected in the classes of readers who process words or nonwords serially, but not in the classes of readers who process words or nonwords in parallel. Accuracy rates and correct reading latencies to words and nonwords of three, four and five letters are presented in Table 5.

Multilevel models were used to test differences in length effects (Snijders & Bosker, 1999). Within a multilevel model, random factors from participants and items can be captured within one model, instead of separate analyses (Quené & van den Bergh, 2004). Each response to an item (Level 1) represents one case, but these cases are nested under individuals (Level 2). These models are equivalent to, for instance, the repeated measures analysis of variance but have more statistical power, because analyses are based on responses to all separate items instead of a mean score per participant per condition.

The analyses were conducted with MLwiN 2.24 (Rasbash, Steele, Browne, & Goldstein, 2008). Separate models for words and nonwords were specified. In each model dummy variables for each length (three, four, or five letters) by class (serial, parallel processors) combination were computed, amounting to a total of six variables. To test the interactions of class and length as well as the main effect of length, length effects were split in two contrasts. These contrasts specified the differences between three versus four and four versus five letter items. The contrasts were tested simultaneously in a multivariate test, using a chi-square statistic with two degrees of freedom (Tabachnick & Fidell, 2001). The main effects of class were tested with a single contrast, resulting in a chi-square statistic with one degree of freedom.

First, a model was specified for accuracy rates. Because accuracy was dummy coded (0 is incorrect, 1 is correct), a logistic regression procedure was used, assuming a binomial distribution rather than the normal distribution assumed for reaction latencies. Mean accuracy rates were high for both words ($M = 0.95$, $SD = 0.07$) and nonwords ($M = 0.92$, $SD = 0.09$). However, serial processors were significantly less accurate than parallel processors for both words, $\chi^2(1) = 117.12$, $p < .001$, and nonwords, $\chi^2(1) = 135.03$, $p < .001$. Length effects were found in the accuracy rates of both classes for both words (serial processors: $\chi^2(2) = 19.72$, $p < .001$; parallel processors: $\chi^2(2) = 19.06$, $p < .001$) and

Table 5

Accuracy Rates and Reading Latencies (and Standard Deviations) for 3-, 4-, and 5-Letter Words and Nonwords in Serial and Parallel Processors

Length	Words				Nonwords			
	Serial processors (N = 37)		Parallel processors (N = 277)		Serial processors (N = 69)		Parallel processors (N = 245)	
	Acc.	RT	Acc.	RT	Acc.	RT	Acc.	RT
3 letters	.89 (.10)	1,160 (459)	.97 (.05)	607 (109)	.88 (.11)	1,206 (436)	.97 (.05)	638 (102)
4 letters	.78 (.17)	1,590 (704)	.95 (.08)	639 (138)	.77 (.16)	1,572 (671)	.93 (.09)	680 (131)
5 letters	.82 (.16)	1,952 (809)	.96 (.06)	669 (159)	.80 (.18)	1,649 (754)	.94 (.08)	702 (138)

Note. Acc. = accuracy; RT = reaction time.

nonwords (serial processors: $\chi^2(2) = 48.40, p < .001$; parallel processors: $\chi^2(2) = 58.15, p < .001$). These length effects did not differ significantly between classes. The effects could mainly be ascribed to three-letter words and nonwords, which were read more accurately than both four- and five-letter items.

The same model was specified for reading latencies. As can be seen in Table 5, large differences are found between classes in mean reading latencies. These differences in mean latencies might affect the interpretation of possible differences in length effects across classes. Significant differences can reflect absolute differences in length effects but might also be merely proportional differences. Because we were interested in relative differences in the effect of length, we controlled for the differences in overall reading latencies by calculating within-subject *z*-scores (Faust, Balota, Spieler, & Ferraro, 1999). The subject's overall mean reading latency was subtracted from every item's reading latency. The difference was divided by the standard deviation of the subject's latency score distribution based on all 90 word and nonword items.

As expected, length effects for words were larger in serial processors than in parallel processors, $\chi^2(2) = 64.15, p < .001$. Unexpectedly, however, a separate test showed that the effect of length was significant in the parallel processors, $\chi^2(2) = 355.95, p < .001$. For nonwords, length effects were also larger in serial processors than in parallel processors, $\chi^2(2) = 32.69, p < .001$. Again, however, a significant length effect was also found for parallel processors, $\chi^2(2) = 283.36, p < .001$.

Two additional analyses were conducted to control for age and neighborhood size, respectively. The classes of word and nonword parallel processors included more of the older children, whereas the majority of the serial processors were children from Grade 2. To determine whether the differences in length effects between classes could be ascribed to age, we conducted the same analyses including only second graders. These children were more equally divided over the classes (words: serial processors $N = 34$, parallel processors $N = 83$; nonwords: serial processors $N = 51$, parallel processors $N = 66$) and did not differ in age (words: 8 years 1 month versus 8 years; nonwords: 8 years 1 month versus 7 years 11 months). Nevertheless, the results in Grade 2 were the same as for the entire group. Length effects were larger in serial than in parallel processors (words: $\chi^2(2) = 35.01, p < .001$; nonwords: $\chi^2(2) = 22.20, p < .001$). Again, length effects were found in both classes of readers for both words (serial processors: $\chi^2(2) = 157.75, p < .001$; parallel processors: $\chi^2(2) = 234.96, p < .001$) and nonwords (serial processors: $\chi^2(2) = 184.90, p < .001$; parallel processors: $\chi^2(2) = 110.00, p < .001$). Thus, differences in the length effects of serial and parallel processors cannot be ascribed to differences in age between the classes.

According to the PDP model, length effects could be ascribed to orthographic neighborhood size (Seidenberg & Plaut, 1998). Therefore, neighborhood size was added to the model for reading latencies as a covariate. Because the distribution of neighborhood size was skewed, a log-transformation was used and neighborhood size was standardized. As a result the estimates for neighborhood size can be interpreted as beta-coefficients. Four dummy variables were specified and added to the models for words and nonwords; the effect of neighborhood size on words and on nonwords in each class separately. The effect of neighborhood size on words was significant only for the parallel processors, $\beta =$

-.06, $\chi^2(1) = 11.19, p < .001$. Words with a larger neighborhood size were read faster than words with a smaller neighborhood size. The effect of neighborhood size on nonwords was significant for both serial processors, $\beta = -.20, \chi^2(1) = 22.75, p < .01$, and parallel processors, $\beta = -.18, \chi^2(1) = 67.62, p < .001$. Nonwords with a larger neighborhood size yielded shorter response latencies than nonwords with a smaller neighborhood size. Although length effects decreased when neighborhood size was controlled for, all length effects remained significant. Thus, length effects in word and nonword reading latencies could not (fully) be ascribed to neighborhood size.

Cross Classification of Classes for Word and Nonword Reading

We combined the classes that were identified in the separate word and nonword models. Interestingly, of the four possible classes, only three classes of readers emerged. The first class consisted of 36 children, who read both words and nonwords serially. These children relied on decoding for both types of letter strings. A second class of 244 children read both words and nonwords in parallel. Finally, 33 children read words in parallel but relied on serial processing for nonwords. Only one child was identified as a serial processor of words but parallel processor of nonwords, indicating that this fourth class of readers did not exist in the data.

Discussion

In the current study we used serial and discrete digit naming to examine serial and parallel processes in word and nonword reading. In line with the results of de Jong (2011), we found that the pattern in the correlations of discrete word reading with serial and discrete digit naming changes over time. From second to fifth grade the relation of discrete word reading with serial digit naming decreased, whereas its relation with discrete digit naming increased. A novel finding was that a similar pattern was found between the formats of digit naming and nonword reading. Regression analyses revealed that from second to fifth grade the amount of unique variance explained by discrete naming increased in both word and nonword reading. Previous studies have also shown that the relations of serial digit naming with word and nonword reading are similar, at least in more transparent orthographies (Greek: Georgiou, Papadopoulos, Fella, & Parrila, 2012; German: Moll, Fussenegger, Willburger, & Landerl, 2009; Dutch: van den Boer et al., 2013). The current results indicate that the development of the relations with both discrete and serial naming over time is similar for words and nonwords.

Next, as predicted, we identified two classes of readers for word reading based on the correlations with serial and discrete digit naming. In line with de Jong (2011), for a large class of readers, single word reading was strongly related to discrete digit naming. For these readers, the process of reading a single word mirrored naming of single overlearned symbols. Words, like digits, were read through parallel retrieval of phonological codes. For a second class of readers, however, single word reading was more strongly related to serial digit naming. For these readers, the process of reading a single word more closely resembled the serial naming of multiple overlearned symbols, suggesting that word reading in this

class relies on a serial process. As argued by de Jong (2011), these results for word reading are compatible with a word-specific view of reading development, as assumed for example in the DRC model (e.g., Coltheart et al., 2001), but cannot be explained within a PDP model (e.g., Plaut et al., 1996).

A novel and unexpected finding was that the same classes were found for nonword reading. Strong correlations between nonword reading and serial digit naming were found for one class of readers, suggesting that nonwords were identified through serial reading processes. For a second and larger class of readers, however, nonword reading related most strongly to discrete digit naming, which suggests that the nonwords were processed in parallel. At first sight, these findings seem to be at odds with both the DRC model (e.g., Coltheart et al., 2001) and the PDP model (e.g., Plaut et al., 1996). For nonwords, both models predict one specific, although different, reading strategy. Whereas nonwords should be processed serially according to the DRC model, the PDP model predicts parallel processing of all letter strings, words and nonwords alike.

Moreover, a clear developmental pattern could be seen, although we were unable to study individuals' stability of class assignment or transition across classes in the current cross-sectional study. First, although in latent class analysis, as opposed to group-wise comparisons, no assumptions have to be made about equal development of all children within an age group (see Bouwmeester & Verkoeijen, 2010), grade level was found to be a good proxy of the class assignments. The classes of serial processors of both words and nonwords consisted mainly of younger children from Grade 2. With just a few exceptions, all the older children in Grades 3 and 5 were able to process the words and nonwords in parallel. These results are in line with de Jong (2011), who identified serial decoders among first and second graders, but not fourth graders. On the other hand, like Ehri and Wilce (1983), we also found parallel processing in young readers with limited reading experience. Even the poorer readers eventually read all words in parallel, since hardly any serial processors were identified past Grade 2.

Second, when class assignments for word and nonword reading were combined, three classes of readers were identified: readers who processed both words and nonwords in parallel, readers who processed only words in parallel, and readers who relied on serial decoding for both words and nonwords. Importantly, the fourth possible group of readers, who process words serially but nonwords in parallel, was not found. Together, these results suggest a developmental path. With increasing reading experience, a shift seems to occur from a serial decoding strategy to identify every letter string toward parallel processing of only words, and later on, even nonwords.

An alternative interpretation of the classes and patterns of correlations in the current study could be the increasing differentiation of abilities over time. In other words, our discrete reading task becomes more strongly related to discrete digit naming because of similar format and task demands. However, if this interpretation is valid, a drop in the relation between serial and discrete digit naming would be expected. Our data do not show a difference in the relation between serial and discrete digit naming across classes of readers: .47 for serial word readers, and .44 for parallel word readers. A similar pattern was found by de Jong (2011), who reported correlations of .50 and .45 for serial and parallel pro-

sors, respectively. In comparison, in that same study the relation between serial and discrete reading dropped from .80 to .32. In addition, increasing differentiation of abilities is most likely a gradual process. In light of a gradual differentiation process, it would not follow that at a certain point in time two classes could be distinguished for whom the tasks either are or are not differentiated. Probably, more than two classes would be found.

To further support our interpretation of the classes of readers, length effects were examined. As predicted, for both words and nonwords we found that length effects were much larger in the classes denoted as serial processors than in the classes of parallel processors. This pattern of results was found both in the entire sample and in a separate analysis of second grade children (i.e., controlling for age differences).

The larger sensitivity to word and nonword length in the class of serial decoders supports our interpretation of the reading strategy used. However, the small length effects in the classes denoted as parallel processors are not in accordance with the general idea that parallel processing of letter strings would result in the absence of a length effect. These findings could imply many different things. Of course, the results could indicate that our interpretation of the difference in reading strategies across the classes is incorrect. For several reasons, however, we think it safe to assume that small length effects can be observed in parallel processors. First, similar small length effects have been regularly reported in advanced adult readers (e.g., Balota, Cortese, Sergent-Marshall, Spieler, & Yap, 2004; Bates, Burani, d'Amico, & Barca, 2001; Ziegler, Perry, Jacobs, & Braun, 2001) and children (Ziegler et al., 2003), all of whom are assumed to use parallel processing. Such small length effects could reflect the involvement of the nonlexical route. Although parallel activation of phonology is the dominant reading strategy, letter strings are simultaneously processed through the nonlexical route. Possibly, the nonlexical route contributed to the identification of at least some of the items. In addition, a small percentage of serial decoders could have been erroneously assigned to the class of parallel processors, which could also result in small length effects. Alternatively, the findings could add to previous indications that length effects cannot be uniformly ascribed to serial decoding of letters into phonological codes (e.g., Risko et al., 2011; van den Boer et al., 2012). In several computational models, length effects are also not ascribed to serial activation of phonology. Within PDP models, for example, length effects are assumed to reflect visual and articulatory factors or neighborhood size (Seidenberg & Plaut, 1998). Alternatively, in more recent connectionist dual process models (CDP⁺: Perry, Ziegler, & Zorzi, 2007; CDP⁺⁺: Perry, Ziegler, & Zorzi, 2010), graphemes are serially connected to the onset, vowel, or coda position in a graphemic buffer. Subsequently, phonology for the input in the graphemic buffer is activated in parallel, either through the lexical route or through a sublexical parallel network of orthographic and phonological units.

In line with this final point, our results do raise the more general question of what is initially processed serially. In line with the DRC framework (e.g., Coltheart et al., 2001; Pritchard, 2012), we have interpreted serial processing as serial activation of phonology through grapheme-phoneme conversion along the nonlexical route. However, serial processing could also occur at the preceding level

of letter identification.² Within the DRC model, as a model of skilled reading, letter features and identities are always identified in parallel. In their work on the causes of letter-by-letter dyslexia, however, Fiset and colleagues (e.g., Fiset, Arguin, Bub, Humphreys, & Riddoch, 2005; Fiset, Arguin, & McCabe, 2006; Fiset, Gosselin, Blais, & Arguin, 2006) highlight that serial reading processes can also occur at the level of letter encoding. When presented with words, readers who suffer from letter-by-letter dyslexia experience an abnormally low signal-to-noise ratio. As a result, these readers present with an impairment at the letter encoding level because visual features of individual letters cannot be registered with enough precision to activate the corresponding letter identities in parallel. Consequently, readers rely on a compensatory sequential letter processing strategy, and focus on each letter separately to achieve the increase in the resolution of the visual system necessary to encode the letter. Possibly, our younger readers, similar to readers who suffer from letter-by-letter dyslexia, were unable to encode the letter strings in parallel and instead processed letters sequentially, irrespective of how phonology was subsequently activated. With increasing reading experience, readers might develop the skills necessary for parallel letter identification, as seen in adults.

This alternative interpretation could account for several of our findings, such as the fact that even among beginning readers, relatively few children were identified as serial processors. It would also be less surprising that similar shifts from serial toward parallel processing were seen in both word and nonword reading. Letter identification should be similar for both types of letter strings. Interestingly, interpreting our results in terms of development in letter processing skills would mean that our findings could be in line with the DRC model. Our idea that nonword phonology could be activated in parallel would be at odds with the DRC model, according to which nonwords are predominately processed through the nonlexical route. If, however, our findings on reading development in children should be interpreted in terms of development in letter identification processes, they could easily be accommodated within the DRC model with the addition of a developmental process in the initial stage of letter identification.

Some of our findings, however, appear difficult to explain through increases in parallel letter encoding, such as the developmental trends indicating parallel processing of words before nonwords. A specific group of children was identified who appeared to process words in parallel, but nonwords serially. If it is letter features and identities that are increasingly processed in parallel, no differences should be expected in the way words and nonwords are processed, given that the initial stage of visual feature and letter encoding is the same for all letter strings. Furthermore, the correlation of word reading with discrete digit naming seems difficult to interpret. This relation was significant for both serial and parallel processors and appeared to increase when words are processed in parallel. Since only a single digit is presented in a discrete naming task, the task cannot reflect parallel identification of multiple items. It could be argued, however, that it is not the number of items that is essential in this relation but rather parallel activation of all the features of an item, be that a single digit or multiple letters. Nevertheless, although visual feature identification could account for some individual differences in discrete digit naming, it is unlikely to account for the relatively high correlation with reading, given the general agreement that discrete digit naming

reflects the retrieval of phonological codes from memory (Bowers & Swanson, 1991; Jones, Branigan, & Kelly, 2009; Logan & Schatschneider, 2014). Taken together, it seems difficult to determine exactly what is initially processed serially. Future studies could help to examine whether it is mainly letters, mainly phonological codes, or both that are increasingly activated in parallel.

Admittedly, the approach taken in the current study adopts assumptions and has limitations that should be mentioned. First, we have to acknowledge that in the current study only short, regular monosyllabic words were studied. The focus on monosyllabic words fits well with the models of the reading system that were studied. Both the DRC (e.g., Coltheart et al., 2001) and PDP (e.g., Plaut et al., 1996) models focus on monosyllabic word reading. The question remains, however, whether the shift from serial toward parallel processing can only be found in short words or could also be seen in longer monosyllabic or in polysyllabic words. In addition, the nonwords in the current study were constructed by interchanging onsets and rhymes of the words. Possibly, nonwords were processed like words, because of their high resemblance to words. Future studies might include multiple sets of nonwords, varying in their similarity to words.

Another limitation lies in the tasks used in the current study. We included only a discrete reading task. Thus, our results cannot be generalized to serial reading tasks. We also made specific choices in the scoring of the discrete naming and reading tasks. The reaction latencies obtained in the naming tasks, which are a measure of time, were converted to fluency scores, a measure of speed. This transformation was chosen to correct for the skewed distributions of reading latencies (Ratcliff, 1993). Our results are not expected to be different, however, if reaction latencies are used, since a high correlation ($r > .80$) was found between fluency scores and reaction latencies for both word and nonword reading. Moreover, the fluency scores as obtained from the discrete word reading task mainly reflect accuracy and automaticity in sublexical and lexical processes, which could also be referred to as reading rate. Our definition therefore differs from fluency measures based on text reading, when for example prosody or comprehension can also be taken into account (e.g., Kuhn, Schwanenflugel, & Meisinger, 2010; Wolf & Katzir-Cohen, 2001). Furthermore, the discrete reading and digit naming task were presented on a computer screen, but the serial naming task was not. However, we do not think this had a major effect on our results. Protopapas, Altani, and Georgiou (2013) administered both serial and discrete naming tasks on a computer and found similar relations with word reading as in the current study.

Finally, we chose to include naming of digits rather than letters. When studying word reading, letter naming might seem the more obvious choice. Digits were chosen, however, because digit names were expected to be even more well known by the children, especially in second grade. In the Netherlands, the names of letters are learned after letter sounds. Digit names are acquired earlier. Moreover, in Dutch, digit names are monosyllabic words, similar to the items in the reading task. However, results are not expected to be different if letters are used. De Jong (2011) presented correlations of discrete word reading with both digit and letter

² We thank an anonymous reviewer for bringing this alternative explanation to our attention.

naming and showed that past Grade 1, relations of letter and digit naming with word reading were found to be almost identical.

Taken together, the results suggest that readers can be sorted into latent classes of serial and parallel processors in reading single monosyllabic words and nonwords based on the relations with serial and discrete digit naming. The different classes were validated by large differences in sensitivity to word and nonword length. Together, the different classes identified suggest a developmental shift from reading all letter strings serially toward parallel processing of words, and later on nonwords. These findings possibly challenge current models of the reading system (e.g., Coltheart et al., 2001; Plaut et al., 1996) and highlight the need for models of the reading system that can accommodate developmental changes from initial serial processing, toward later parallel processing of all letter strings.

References

Aaron, P. G., Joshi, R. M., Ayotollah, M., Ellsberry, A., Henderson, J., & Lindsey, K. (1999). Decoding and sight-word naming: Are they independent components of word recognition skill? *Reading and Writing*, 11, 89–127. doi:10.1023/A:1008088618970

Balota, D. A., Cortese, M. J., Sergent-Marshall, S. D., Spieler, D. H., & Yap, M. J. (2004). Visual word recognition of single-syllable words. *Journal of Experimental Psychology: General*, 133, 283–316. doi:10.1037/0096-3445.133.2.283

Bates, E., Burani, C., d'Amico, S., & Barca, L. (2001). Word reading and picture naming in Italian. *Memory & Cognition*, 29, 986–999. doi:10.3758/BF03195761

Bouwmeester, S., & Verkoeijen, P. P. J. L. (2010). Latent variable modeling of cognitive processes in true and false recognition of words: A developmental perspective. *Journal of Experimental Psychology: General*, 139, 365–381. doi:10.1037/a0019301

Bowers, P. G., & Swanson, L. B. (1991). Naming speed deficits in reading disability: Multiple measures of a singular process. *Journal of Experimental Child Psychology*, 51, 195–219. doi:10.1016/0022-0965(91)90032-N

Brus, B., & Voeten, B. (1995). *Eén minuut test vorm A en B. Verantwoording en handleiding* [One-Minute Test, Forms A and B: Justification and manual]. Lisse, the Netherlands: Swets & Zeitlinger.

Celeux, G., & Soromenho, G. (1996). An entropy criterion for assessing the number of clusters in a mixture model. *Journal of Classification*, 13, 195–212. doi:10.1007/BF01246098

Coltheart, M., Rastle, K., Perry, C., Langdon, R., & Ziegler, J. (2001). DRC: A dual route cascaded model of visual word recognition and reading aloud. *Psychological Review*, 108, 204–256. doi:10.1037/0033-295X.108.1.204

de Jong, P. F. (2011). What discrete and serial rapid automatized naming can reveal about reading. *Scientific Studies of Reading*, 15, 314–337. doi:10.1080/10888438.2010.485624

Denckla, M. B., & Rudel, R. (1976). Naming of object-drawings by dyslexic and other learning disabled children. *Brain and Language*, 3, 1–15. doi:10.1016/0093-934X(76)90001-8

Ehri, L. C. (2005). Learning to read words: Theory, findings, and issues. *Scientific Studies of Reading*, 9, 167–188. doi:10.1207/s1532799xssr0902_4

Ehri, L. C., & Wilce, L. S. (1983). Development of word identification speed in skilled and less skilled beginning readers. *Journal of Educational Psychology*, 75, 3–18. doi:10.1037/0022-0663.75.1.3

Faust, M. E., Balota, D. A., Spieler, D. H., & Ferraro, F. R. (1999). Individual differences in information-processing rate and amount: Implications for group differences in response latency. *Psychological Bulletin*, 125, 777–799. doi:10.1037/0033-2909.125.6.777

Fiset, D., Arguin, M., Bub, D., Humphreys, G. W., & Riddoch, M. J. (2005). How to make the word-length effect disappear in letter-by-letter dyslexia: Implications for an account of the disorder. *Psychological Science*, 16, 535–541. doi:10.1111/j.0956-7976.2005.01571.x

Fiset, D., Arguin, M., & McCabe, E. (2006). The breakdown of parallel letter processing in letter-by-letter dyslexia. *Cognitive Neuropsychology*, 23, 240–260. doi:10.1080/02643290442000437

Fiset, D., Gosselin, F., Blais, C., & Arguin, M. (2006). Inducing letter-by-letter dyslexia in normal readers. *Journal of Cognitive Neuroscience*, 18, 1466–1476. doi:10.1162/jocn.2006.18.9.1466

Forster, K. I., & Chambers, S. M. (1973). Lexical access and naming time. *Journal of Verbal Learning & Verbal Behavior*, 12, 627–635. doi:10.1016/S0022-5371(73)80042-8

Frederiksen, J. R., & Kroll, J. E. (1976). Spelling and sound: Approaches to the internal lexicon. *Journal of Experimental Psychology: Human Perception and Performance*, 2, 361–379. doi:10.1037/0096-1523.2.3.361

Georgiou, G. K., Papadopoulos, T. C., Fella, A., & Parrila, R. (2012). Rapid naming speed components and reading development in a consistent orthography. *Journal of Experimental Child Psychology*, 112, 1–17. doi:10.1016/j.jecp.2011.11.006

Jackson, N. E., & Coltheart, M. (2001). *Routes to reading success and failure: Toward an integrated cognitive psychology of atypical reading*. New York, NY: Taylor & Francis.

Jones, M. W., Branigan, H. P., & Kelly, M. L. (2009). Dyslexic and nondyslexic reading fluency: Rapid automatized naming and the importance of continuous lists. *Psychonomic Bulletin & Review*, 16, 567–572. doi:10.3758/PBR.16.3.567

Kuhn, M. R., Schwanenflugel, P. J., & Meisinger, E. B. (2010). Aligning theory and assessment of reading fluency: Automaticity, prosody, and definitions of fluency. *Reading Research Quarterly*, 45, 230–251. doi:10.1598/RRQ.45.2.4

Logan, J. A. R., & Schatschneider, C. (2014). Component processes in reading: Shared and unique variance in serial and isolated naming speed. *Reading and Writing*, 27, 905–922. doi:10.1007/s11145-013-9475-y

Lubke, G. H., & Muthén, B. O. (2005). Investigating population heterogeneity with factor mixture models. *Psychological Methods*, 10, 21–39. doi:10.1037/1082-989X.10.1.21

Marinus, E., & de Jong, P. F. (2010). Variability in the word-reading performance of dyslexic readers: Effects of letter length, phoneme length and digraph presence. *Cortex*, 46, 1259–1271. doi:10.1016/j.cortex.2010.06.005

Moll, K., Fussenegger, B., Willburger, E., & Landerl, K. (2009). RAN is not a measure of orthographic processing: Evidence from the asymmetric German orthography. *Scientific Studies of Reading*, 13, 1–25. doi:10.1080/10888430802631684

Muthén, L. K., & Muthén, B. O. (2009). *Mplus (Version 5.21)* [Computer software]. Los Angeles, CA: Muthén & Muthén.

Nylund, K. L., Asparouhov, T., & Muthén, B. O. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study. *Structural Equation Modeling*, 14, 535–569. doi:10.1080/10705510701575396

Perry, C., Ziegler, J. C., & Zorzi, M. (2007). Nested incremental modeling in the development of computational theories: The CDP model of reading aloud. *Psychological Review*, 114, 273–315. doi:10.1037/0033-295X.114.2.273

Perry, C., Ziegler, J. C., & Zorzi, M. (2010). Beyond single syllables: Large-scale modeling of reading aloud with the Connectionist Dual Process (CDP⁺⁺) model. *Cognitive Psychology*, 61, 106–151. doi:10.1016/j.cogpsych.2010.04.001

Plaut, D. C. (1999). A connectionist approach to word reading and acquired dyslexia: Extension to sequential processing. *Cognitive Science*, 23, 543–568. doi:10.1207/s15516709cog2304_7

Plaut, D. C., McClelland, J. L., Seidenberg, M. S., & Patterson, K. (1996). Understanding normal and impaired word reading: Computational principles in quasi-regular domains. *Psychological Review*, 103, 56–115. doi:10.1037/0033-295X.103.1.56

Pritchard, S. C. (2012). *Incorporating learning mechanisms into the dual-route cascaded (DRC) model of reading aloud and word recognition*. (Unpublished doctoral dissertation). Macquarie University, Sydney, Australia.

Protopapas, A., Altani, A., & Georgiou, G. K. (2013). Development of serial processing in reading and rapid naming. *Journal of Experimental Child Psychology*, 116, 914–929. doi:10.1016/j.jecp.2013.08.004

Quené, H., & van den Bergh, H. (2004). On multi-level modeling of data from repeated measures designs: A tutorial. *Speech Communication*, 43, 103–121. doi:10.1016/j.specom.2004.02.004

Rasbash, J., Steele, F., Browne, W. J., & Goldstein, H. (2008). *A user's guide to MLwiN* (Version 2.10). Bristol, England: University of Bristol, Centre for Multilevel Modelling.

Ratcliff, R. (1993). Methods for dealing with reaction time outliers. *Psychological Bulletin*, 114, 510–532. doi:10.1037/0033-2909.114.3.510

Risko, E. F., Lanthier, S. N., & Besner, D. (2011). Basic processes in reading: The effect of interletter spacing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 37, 1449–1457. doi:10.1037/a0024332

Schneider, W., Eschman, A., & Zuccolotto, A. (2002). *E-prime: User's guide*. Pittsburgh, PA: Psychology Software.

Schrooten, W., & Vermeer, A. (1994). *Woorden in het basisonderwijs: 15.000 woorden aangeboden aan leerlingen*. [Words in primary education: 15,000 words presented to students]. Tilburg, the Netherlands: Tilburg University Press.

Seidenberg, M. S., & McClelland, J. L. (1989). A distributed, developmental model of word recognition and naming. *Psychological Review*, 96, 523–568. doi:10.1037/0033-295X.96.4.523

Seidenberg, M. S., & Plaut, D. C. (1998). Evaluating word-reading models at the item level: Matching the grain of theory and data. *Psychological Science*, 9, 234–237. doi:10.1111/1467-9280.00046

Share, D. L. (1995). Phonological recoding and self-teaching: Sine qua non of reading acquisition. *Cognition*, 55, 151–218. doi:10.1016/0010-0277(94)00645-2

Share, D. L. (1999). Phonological recoding and orthographic learning: A direct test of the self-teaching hypothesis. *Journal of Experimental Child Psychology*, 72, 95–129. doi:10.1006/jecp.1998.2481

Share, D. L. (2008). On the anglocentricities of current reading research and practice: The perils of overreliance on an “outlier” orthography. *Psychological Bulletin*, 134, 584–615. doi:10.1037/0033-2909.134.4.584

Snijders, T., & Bosker, R. (1999). *Multilevel modeling: An introduction to basic and advanced multilevel modeling*. London, England: Sage.

Spinelli, D., de Luca, M., Filippo, G. D., Mancini, M., Martelli, M., & Zuccolotti, P. (2005). Length effect in word naming in reading: Role of reading experience and reading deficit in Italian readers. *Developmental Neuropsychology*, 27, 217–235. doi:10.1207/s15326942dn2702_2

Tabachnick, B. G., & Fidell, L. S. (2001). *Using multivariate statistics* (4th ed.). Boston, MA: Allyn & Bacon.

van den Boer, M., de Jong, P. F., & Haentjens-van Meeteren, M. M. (2012). Lexical decision in children: Sublexical processing or lexical search? *The Quarterly Journal of Experimental Psychology*, 65, 1214–1228. doi:10.1080/17470218.2011.652136

van den Boer, M., de Jong, P. F., & Haentjens-van Meeteren, M. M. (2013). Modeling the length effect: Specifying the relation with visual and phonological correlates of reading. *Scientific Studies of Reading*, 17, 243–256. doi:10.1080/10888438.2012.683222

Van den Bos, K. P., Zijlstra, B. J. H., & Van den Broeck, W. (2003). Specific relations between alphanumeric-naming speed and reading speeds of monosyllabic and multisyllabic words. *Applied Psycholinguistics*, 24, 407–430. doi:10.1017/S0142716403000213

Weekes, B. S. (1997). Differential effects of number of letters on word and nonword naming latency. *The Quarterly Journal of Experimental Psychology A: Human Experimental Psychology*, 50, 439–456. doi:10.1080/713755710

Wolf, M., & Katzir-Cohen, T. (2001). Reading fluency and its intervention. *Scientific Studies of Reading*, 5, 211–239. doi:10.1207/S1532799XSSR0503_2

Ziegler, J. C., Perry, C., Jacobs, A. M., & Braun, M. (2001). Identical words are read differently in different languages. *Psychological Science*, 12, 379–384. doi:10.1111/1467-9280.00370

Ziegler, J. C., Perry, C., Ma-Wyatt, A., Ladner, D., & Schulte-Körne, G. (2003). Developmental dyslexia in different languages: Language specific or universal? *Journal of Experimental Child Psychology*, 86, 169–193. doi:10.1016/S0022-0965(03)00139-5

Zuccolotti, P., de Luca, M., di Pace, E., Gasperini, F., Judica, A., & Spinelli, D. (2005). Word length effect in early reading and in developmental dyslexia. *Brain and Language*, 93, 369–373. doi:10.1016/j.bandl.2004.10.010

Received May 24, 2013

Revision received March 6, 2014

Accepted April 20, 2014 ■

Classmate Characteristics and Student Achievement in 33 Countries: Classmates' Past Achievement, Family Socioeconomic Status, Educational Resources, and Attitudes Toward Reading

Ming Ming Chiu
University at Buffalo

Bonnie Wing-Yin Chow
City University of Hong Kong

Classmates can influence a student's academic achievement through immediate interactions (e.g., academic help, positive attitudes toward reading) or by sharing tangible or intangible family resources (books, stories of foreign travel). Multilevel analysis of 141,019 fourth-grade students' reading achievements in 33 countries showed that classmates' family factors (parent socioeconomic status [SES], home educational resources) were more strongly related to a student's reading achievement than were classmates' characteristics (parent ratings of past literacy skills, attitudes toward reading). However, these classmate links to reading achievement differed across students (e.g., high-SES classmates benefited high-SES students more than low-SES students). Also, links between classmates' past reading achievement and a student's current reading achievement were stronger in countries that were richer, were more collectivist, or avoided uncertainty less. These findings show how an ecological model of family and classmate microsystems, classmate family mesosystem, and country macrosystem can help provide a comprehensive account of children's academic achievement.

Keywords: Bronfenbrenner, ecological system theory, classmates, literacy, cross-cultural study

Classmates play a significant role in children's behaviors and academic achievement (Opdenakker & Van Damme, 2001). Specifically, a student with higher achieving classmates often shows higher academic achievement than does a student with lower achieving classmates (Kang, 2007; Zimmer & Toma, 2000). These classmates' influences contribute to a complex environmental system in which various levels of factors interact. However, past studies have not explicated the mechanisms by which classmates affect a student's learning. Furthermore, they have not tested whether these links are universal or whether they differ across countries. Therefore, the present study helps fill these research gaps by testing a model of how family, classmate, and country characteristics are related to academic achievement among elementary school children. The proposed model was supported through the reading tests and associated survey data of a representative sample of 141,019 fourth graders in 33 countries.

Environmental Influences on Academic Achievement

Children develop within a complex environment that consists of multiple levels of surrounding contexts, according to Bronfen-

brenner's (2005) ecological system theory. The immediate contexts (*microsystems*; e.g., family, classroom), the relationships between microsystems (*mesosystem*; e.g., classmates' families) and the broader country resources (*macrosystem*; e.g., economy, cultural values) can contribute to student learning (see Figure 1). While some relationships are universal, others differ across countries.

Microsystem: Family

Students in higher socioeconomic status (SES) families (or high-SES students) have higher academic achievement, even after accounting for genetic factors (e.g., Walker, Petrill, & Plomin, 2005). Families can use their financial, human, social, and cultural capital to give their children learning opportunities (Chiu, 2013). Specifically, families with more money (*financial capital*) can buy more educational resources (books, calculators, and so on) to create a richer learning environment (Chiu, 2010). Furthermore, high-SES students often spend more time with their parents (due to fewer competing siblings, less parent time on housework, and multitasking parents), so they can capitalize more on their parents' human, social, and cultural capital. Families with more education, knowledge, and skills (*human capital*) often create better learning environments for their children, foster better attitudes toward reading, and teach them more skills than other families can (Davalos, Chavez, & Guardiola, 2005; Willms, 1999).

High-SES families also often have substantial social or cultural capital. High-SES families typically have large social networks of relatives, friends, and acquaintances with skills or resources (*social capital*) that can help their children learn (Chiu, 2013). Likewise, high-SES families often have cultural possessions or experiences (*cultural capital*) that can help their children learn their

This article was published Online First June 2, 2014.

Ming Ming Chiu, Department of Learning and Instruction, University at Buffalo; Bonnie Wing-Yin Chow, Department of Applied Social Studies, City University of Hong Kong.

We appreciate the research assistance of Yik Ting Choi. This research was partially funded by a grant from the Spencer Foundation.

Correspondence concerning this article should be addressed to Ming Ming Chiu, who is now at the Department of Educational Studies, Purdue University, 100 North University Street, West Lafayette, IN 47907. E-mail: mingmingchiu@gmail.com

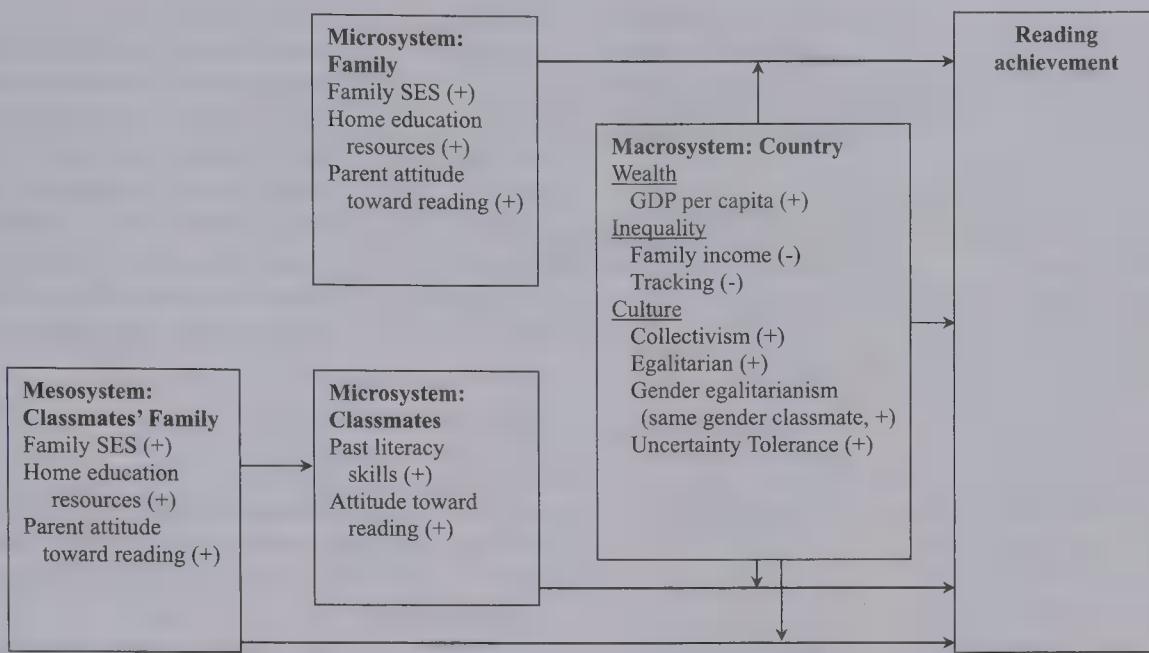


Figure 1. Model of classmates' effects on students' reading achievement (control variables are not shown). SES = socioeconomic status; GDP = gross domestic product.

society's cultural knowledge, skills, and values to adapt to their school culture (Chiu & Chow, 2010).

In short, high-SES students have more financial, human, social, or cultural capital. Using their greater capital, higher SES students can better understand others' expectations, behave properly at school, have closer relationships with teachers and classmates, and learn more in school than lower SES students do.

Microsystem: Classmates

Children spend a large proportion of time at school and interact regularly with their classmates. If a student has a larger social network of classmates or stronger relationships with them, he or she has more social capital on which to capitalize and learn more (Pong, 1997, 1998). Moreover, high-SES students might use their superior resources and skills to benefit more from classmates' resources and experiences.

Classmate benefits. Classmates can help a student learn both directly and indirectly (Skibbe, Phillips, Day, Brophy-Herb, & Connor, 2012). Classmates can directly help a student by sharing information or evaluations. For example, a classmate can explain the meaning of a vocabulary word (Murphy, 1994). Also, when a student proposes an idea, a classmate can recognize its validity and further justify it (Chiu, 2008) or identify its flaws and correct it (Chiu & Khoo, 2003). Thus, classmates can provide information, validate correct ideas, or recognize flaws to help a student learn.

Classmates can also help a student learn indirectly through motivation and norms. They can motivate a student to enjoy learning, which helps them exert effort and persevere when facing setbacks (Chiu & McBride-Chang, 2006). For example, a classmate can show enthusiasm for a storybook character, which can entice a student to study together (Edmunds & Bauserman, 2006; Guthrie, Klauda, & Morrison, 2012; Skibbe et al., 2012). When a student performs poorly on a reading test, a classmate can provide emotional support and encourage further study so that the student can do better on the next test (Guthrie et al., 2012; Skibbe et al.,

2012). Hence, classmates can motivate a student via greater enjoyment, study time, and perseverance.

In addition to motivation, classmates can help create and maintain norms of attitude, behavior, and achievement, both among friends and within the classroom. Classmates can articulate and model positive academic attitudes, behave within discipline norms, study hard, and perform well on tests, essays, and other academic measures. Together, classmates can cultivate a culture of positive attitudes toward reading in which to immerse a student (Johnson & Johnson, 1999). Supported by these positive attitudes, classmates can model appropriate classroom behavior (e.g., raise hands to answer teacher questions, rather than interrupting) and encourage a student to behave accordingly (Lewis, 2001; Ma & Willms, 2004). Buttressed by these attitudes and behaviors, classmates are more likely to have higher reading achievement, which raises a student's academic expectations (Baker & Wigfield, 1999).

Unequal benefits. However, classmates do not benefit each student equally. While high-SES students can share some resources publicly, they may share other resources privately. Privately shared resources are an incentive for classmates to try to befriend high-SES students rather than low-SES students. As a result, high-SES students often attract more and build stronger relationships with them. Hence, high-SES students might build and capitalize on a stronger network of classmates to learn more (Ryan, 2001).

As higher SES students may have more capital, higher status, and better interaction skills than other students, they can entice high-SES classmates into their network and build closer relationships. As noted earlier, a high-SES student often learns strong social skills from family members and can use them to build closer relationships with classmates (Chiu & Chow, 2010). As people prefer to interact with those who are similar to them (common language, similar activities, similar preferences, and so on; *homophily bias*, McPherson, Smith-Lovin, & Cook, 2001; or *assortativeness*, Kindermann, 2007), high-SES classmates often share more

common experiences with high-SES students and prefer to interact with them, compared with low-SES students (Chiu, in press). Hence, high-SES students often have stronger social networks with high-SES classmates and capitalize on them to learn more than other students (Crosnoe, 2004).

Moreover, not all interactions with classmates necessarily enhance a student's learning. Social capital is positively related with academic achievement among majority populations (e.g., native-born students; Coleman, 1994) but not among minority or marginalized populations (e.g., ethnic minority), which often have less social capital (negative or nonsignificant relations; Ream, 2003). As students in marginalized groups often devote extensive effort to develop tight ties within their lower social capital group, they have less time and effort to develop ties with higher social capital groups (Ream, 2003). Therefore, students in marginalized groups often draw upon less social capital (or even suffer negative effects through poor academic attitudes, behaviors, or norms; Ream, 2003). Together, these studies suggest that the link between social capital and academic achievement depends on the student's own resources, attitudes, and skills.

Also, students with more of a specific resource, attitude, or skill might benefit more from that of their classmates. These dimension-specific benefits might operate through selective trading, norm establishment/maintenance, or skill practice. For example, a student with more books (or other educational resources) might trade or lend them to classmates for other books (Mankiw, 2011). In this way, richer students with more resources have greater access than poorer students to classmates' resources and can benefit more from them. Classmate-established norms (e.g., lunch chats about story books) might help a student who enjoys reading to learn more than a student who does not enjoy reading (Chiu & McBride-Chang, 2006). Also, classmates with a skill (e.g., read English) might trade more information with a student with some of that skill (knows some words) than with another student without any of it (knows no words). Or classmates with that skill can help a student with some of that skill practice and thereby improve it more than someone who lacks the skill entirely.

In short, classmates are part of a student's social capital. Classmates can help a student learn directly (sharing information, evaluations) or indirectly (motivation, norms). Moreover, students in higher SES families or in the majority population often benefit more from classmate resources and experiences than other students. Also, students with more of a resource, skill, or attitude might benefit more from that of their classmates via selective trading, participation within the norm, or skill practice.

Mesosystem: Classmates' Families

The mesosystem of students and classmates' families link the family and classmate microsystems. In addition to interactions with classmates, their family capital can help a student learn more (Caldas & Bankston, 1997; Chiu & Zeng, 2008; Sallee & Tierney, 2007). A student can benefit from interactions with classmates' families (e.g., chatting with a classmate's mom about a mayoral election), from their home resources (e.g., borrow poetry book), from their school contributions (e.g., guest talks), or from their interactions with school staff (e.g., active parent-teacher organization). Thus, the human, financial, cultural, or social capital of classmates' families might influence a student's learning.

However, classmates' families do not necessarily help all students equally. As noted previously, high-SES students tend to have stronger social networks of classmates, especially high-SES classmates (Ryan, 2001). Thus, they are more likely to capitalize on their classmates' family capital. Furthermore, marginalized families often have less family capital, so marginalized students with close ties to similar classmates might benefit less from their classmates' family members (or even suffer negative effects through poor academic attitudes, behaviors, or norms). Thus, high-SES students might benefit more from classmates' family capital.

Macrosystem: Country

Classmate effects may differ across countries due to economic conditions or cultural values. In addition to their potential effect on student achievement, a country's economy or cultural values might moderate classmate or classmate family influence on academic achievement.

Economy. Richer countries (e.g., Switzerland) often have more public resources that can raise student achievement or the influence of high-achieving classmates. As richer countries often provide more public resources (e.g., public libraries, museums) or better education (e.g., certified English teachers; Baker, Goesling, & Letendre, 2002), students in richer countries often capitalize on these opportunities (e.g., reading more library books) to learn more. In addition, high-achieving classmates in richer countries might use these extra public resources not available in poorer countries to help a student learn more (e.g., explaining the word *newt* with a Wikipedia photo; Chiu, 2007). In richer countries, students might learn more or might be influenced more by high-achieving classmates.

In addition to the amount of resources, the distribution of resources within a country might affect student achievement or classmate influences. Greater household income inequality within a country might reduce student achievement through diminishing marginal returns or homophily bias. Consider a student with two dictionaries. She keeps the first one on her desk and uses it often when she reads. In contrast, the second one sits on a bookcase (unless the first one is lost or is being used). This lower value of the second dictionary (or, more generally, additional resources of the same type) is *diminishing marginal returns* (Chiu & Khoo, 2005). Applied to inequality, a poor student (with one book) likely learns more from an extra book than a rich student (with a hundred books) would. In countries with greater household income equality (such as Norway), poorer students have more resources and might benefit more from them, compared with richer students; resulting in higher achievement overall; hence, students in countries with greater equality might show greater achievement than those in countries with less equality (e.g., Colombia). In a similar vein, viewing students as resources, clustering high-achieving students together in a few classes (tracking) or schools (banding) away from lower achieving students can reduce the latter's access to high-achieving students and reduce overall achievement (Chiu & Khoo, 2005).

Greater equality might also increase overall student achievement through people's homophily bias toward others of similar SES (Kindermann, 2007; McPherson et al., 2001). As a result, students in more equal countries might interact, cooperate, and share resources more often, resulting in higher achievement over-

all. Moreover, increased interactions might increase classmate effects in more equal countries. Meanwhile, clustering high-achieving students together might increase interactions among them due to homophily bias, which can mitigate the negative effects of unequal access and diminishing marginal returns.

Cultural values. Apart from economies, cultural values differ across countries (LeTendre, Hofer, & Shimizu, 2003). Countries differ according to their degree of status hierarchy and obedience to authority versus equality (hierarchical vs. egalitarian; or *power distance*), favoring group interests versus individual interests (collectivism vs. individualism), emphasizing rigidly defined versus flexible gender roles (masculine vs. feminine), and tolerance of risk (uncertainty avoidance vs. uncertainty tolerance; Hofstede, 2003). Past studies have shown that these cultural values have no direct relationship with academic achievement, but they moderate the links between other factors (e.g., self-concept) and academic achievement (Chiu & Klassen, 2009). Consider four hypotheses regarding the moderating effects of cultural influences.

First, collectivistic societies value group interests over individual interests (House, Hanges, Javidan, Dorfman, & Gupta, 2004). In collectivist societies (e.g., Argentina), classmates pay greater attention to one another's preferences (e.g., favorite book), talk with one another more often, and conform more closely to group norms (e.g., listening, turn-taking, politeness) than in individualistic societies (e.g., Russia; Chiu & Chen, *in press-a*). Thus, classmates are more likely to influence a student's learning in collectivist countries than in individualistic countries (Chiu & Chen, *in press-b*). For example, classmates' metacognitive strategies (e.g., set goals, evaluate progress) are more strongly linked to a student's reading achievement in collectivist countries than in individualistic countries (Chiu, Chow, & McBride-Chang, 2007). Hence, classmate characteristics might have stronger links to a student's academic attitudes, behaviors, or achievements in collectivist countries than in individualistic countries.

Second, people in egalitarian countries (e.g., Italy) value equal status and equal opportunity more than those in hierarchical countries do (e.g., New Zealand; Hofstede, 2003). In egalitarian countries, high- and low-status students often attend the same schools, and teachers and classmates tend to treat them similarly, so their learning experiences tend to be more similar than those in hierarchical countries (Chiu & Zeng, 2008). As students who perceive one another as more similar (e.g., more equal) are more likely to interact with one another due to homophily bias, students in egalitarian countries are more likely than those in hierarchical countries to interact with one another (McPherson et al., 2001). Moreover, high-achieving students are more likely to interact with high-achieving students than low-achieving students, so high-achieving classmates might influence a high-achieving student more than a low-achieving student. As noted earlier, high-SES students are more likely to interact with and influence high-SES students than low-SES students. Hence, the effects of classmate characteristics on a student's attitudes, behaviors, or achievements might be stronger both in more egalitarian countries and on students with similar characteristics.

Third, students in more feminine or gender egalitarian countries (e.g., Canada) view gender roles as flexible, whereas those in masculine countries have rigidly defined gender roles (e.g., Kuwait; House et al., 2004). With different gender role expec-

tations, boys and girls might attend different schools, receive different treatment from their parents or teachers, and experience schooling differently (Chiu & Chow, 2010). As a result, girls might interact with girls more often, and boys might interact with boys more often. With different expectations and experiences, boys and girls might prefer to interact with classmates of the same gender rather than different genders due to homophily bias, resulting in weaker social relationships and less influence between boys and girls (McPherson et al., 2001). Then, classmates of the same gender might have greater influence than those of the opposite gender on a student's attitudes, behaviors, or achievements.

Last, students in cultures with greater uncertainty avoidance (e.g., Iran) have lower tolerance of risk. As they prefer familiar people, surroundings, and values, they typically prefer interacting with family and relatives that they have known for most of their lives rather than with relatively new classmates (Hofstede, 2003). As a result, students in these cultures are less likely than students in uncertainty-tolerant cultures (e.g., Netherlands) to spend much time engaging with their classmates. Furthermore, students with high uncertainty avoidance are more likely to value their family's attitudes toward schools (e.g., utility of schooling to future career), follow their siblings' behaviors (e.g., hours of study), and model their academic expectations on those of their siblings (e.g., which grades are acceptable), rather than those of their classmates. As a result, classmates might have less influence on students in cultures with greater uncertainty avoidance than on those in cultures with less uncertainty avoidance. Hence, the effects of classmate characteristics on a student's attitudes, behaviors, or achievements might be weaker in countries with greater uncertainty avoidance than in other countries.

Present Study

The present study tests a model of classmate characteristics and academic achievement among elementary school children in 33 countries (see Table 1). We asked two research questions. First, are classmates' characteristics (including family SES, home literacy resources, attitude toward reading, and past reading achievement) related to student reading achievement? We hypothesized that when classmates have higher family SES, better home literacy resources, better attitudes toward reading, or higher reading performance, students on average have higher reading achievement.

Second, do these links among classmates' factors, classmate family factors, and students' reading achievement differ across countries with different economic characteristics or cultural values? We expected links between classmates' characteristics and a student's reading achievement to be stronger in countries that are richer, more equal, more collectivist, or cluster students together by past achievement (tracking or banding). We expected same-gender classmate characteristics to be stronger in masculine countries, whose gender roles are more rigid. Last, we expected links between classmates' characteristics and a student's reading achievement to be weaker in uncertainty avoidance countries. To account for the possibility that student achievement is related to classmate achievement simply because students of similar past achievement attend class together, we included student past

Table 1
Theoretical Hypotheses

Explanatory variable	Hypothesized outcome	Supported?
Greater classmate characteristics	Higher	
Family socioeconomic status	Student reading achievement	Yes
Educational resources at home	Student reading achievement	Yes
Parent attitude toward reading	Student reading achievement	Yes
Attitude toward reading	Student reading achievement	Yes
Past reading achievement	Student reading achievement	Yes
Greater country characteristics	Stronger	
Wealth	Classmate influence	Yes
Economic equality	Classmate influence	
Clustering students by ability	Classmate influence	
Collectivism	Classmate influence	
Egalitarian	Classmate influence	
Gender egalitarianism	Same gender classmate influence	Yes
Uncertainty tolerance	Classmate influence	Yes

achievement (specifically, parent rating of past literacy skills) as a control variable.

Method

Data

The International Association for the Evaluation of Education Achievement's Progress in International Reading Literacy Study (IEA-PIRLS) assessed 141,019 fourth-grade students' reading achievement and asked students and principals to complete questionnaires related to their perceptions of themselves and their immediate environments (Martin, Mullis, & Kennedy, 2003). Students completed an 80-min assessment booklet and then a 15- to 30-min questionnaire. We also used economic data (World Bank, 2002) and cultural values data (House et al., 2004).

This sample included a variety of countries, ranging from poor, unequal, collectivist nations (e.g., Colombia) to rich, relatively equal, individualistic ones (e.g., Norway). The regions and countries that participated were Argentina, Belize, Bulgaria, Canada (across Alberta, British Columbia, Nova Scotia, Ontario, and Quebec), Colombia, Cyprus, Czech Republic, France, Germany, Greece, Hong Kong, Hungary, Iceland, Iran, Israel, Italy, Kuwait, Latvia, Lithuania, Moldova, Morocco, the Netherlands, New Zealand, Norway, Romania, Russian Federation, Singapore, Slovak Republic, Slovenia, Sweden, Turkey, Macedonia, England, Scotland, and the United States. As all responses to home questionnaires were missing in the Morocco and U.S. data, these two countries were not included in our analysis.

Methodological Design

Investigating the relationships between classmate characteristics and reading achievement across countries requires representative sampling, precise tests and questionnaires, and suitable statistical models. In each country, IEA chose about 150 representative schools based on neighborhood SES and student intake and sampled one or two fourth-grade classes from each school (stratified sampling), resulting in a sample size of about 4,000 students per

country or region (Martin et al., 2003). Students who had intellectual disabilities, refused to take the exam, could not physically take it, or did not understand the test language altogether accounted for less than 4% of the original sample. With suitable weights, IEA created representative samples of each country's schools and fourth-grade students.

Students received subtests (overlapping subsets of all multiple choice and open-ended questions) for wider coverage of reading skills while reducing student fatigue and learning during the test (a balanced incomplete block test; Baker & Kim, 2004). A graded-response Rasch model of these subtests measured the difficulty of each test item to estimate each student's reading competence more precisely (Baker & Kim, 2004).

To reduce measurement error, researchers used several questionnaire items for each theoretical construct (e.g., SES) to create an index via a graded response Rasch model (Warm, 1989). The multigroup graded-response Rasch models for each item in each country yielded similar parameters, indicating measurement equivalence across countries (Martin et al., 2003; May, 2006). (Unlike factor analysis, a multigroup graded-response Rasch model has two advantages: it requires only one invariant anchor item across countries and models heterogeneous use of the ordinal rating scale; Rossi, Gilula, & Allenby 2001). Still, the graded-response Rasch model assumes that the relationships between the items, and the construct are the same across countries. Other studies also have shown consistent questionnaire responses and participant understandings across countries (Brown, Micklewright, Schnepf, & Waldmann, 2007; Martin et al., 2003; Schulz, 2003). To estimate reliability, the graded-response Rasch models included computations of the information function (Baker & Kim, 2004); when the information function is greater, there is more information, smaller standard errors, more precision, and greater reliability (see Table 2 for the reliability of each index). PIRLS standardized the test scores to a mean of zero (and standard deviation of 1) for all data from all countries to facilitate identification of scores above and below the overall mean.

Missing questionnaire response data (8%) can reduce estimation efficiency, complicate data analyses, and bias results. Markov chain Monte Carlo multiple imputation addresses these missing

Table 2
Summary Statistics and Variable Descriptions (N = 141,019)

Variable	M	SD	Description
Reading achievement	503.14	95.07	Reading scores estimated by the graded-response Rasch models were calibrated to $M = 500$ and $SD = 100$ (Martin, Mullis, & Kennedy, 2003). Min = 9.73; Max = 820.63.
Explanatory variables			
Student's parent rating of past literacy skills ^a	0.00	1.00	Index of literacy skills at start of 1st grade: Recognize most of the alphabet letters; read some words; read sentences; write letters of the alphabet; write some words. Response choices = <i>not at all, not very well, moderately well, very well</i> . Min = -2.28; Max = 2.02. Reliability = .95.
Ecological			
Country variables			
GDP per capita	15,925	8,251	Gross domestic product per person (World Bank, 2002). Min = US\$2,085; Max = US\$27,060.
Log GDP per capita	9.49	0.66	World Bank, 2002. This fit the data better than GDP per capita. Min = 7.64; Max = 10.21.
Gini index	35.91	8.03	Scores range from 0 (<i>perfect equality</i> ; <i>same incomes for all</i>) to 100 (<i>perfect inequality</i> ; <i>one person has all the income</i>). (World Bank, 2002) Min = 25; Max = 57.6.
Power distance	2.72	0.33	Multigroup graded-response Rasch-based index (Warm, 1989) of 7-point Likert responses to five questions (See Appendix A; House et al., 2004). Min = 2.19; Max = 3.53; Reliability = .88.
In-group collectivism	4.63	0.46	Multigroup graded-response Rasch-based index (Warm, 1989) of 7-point Likert responses to four questions (See Appendix A; House et al., 2004). Min = 3.98; Max = 5.62; Reliability = .95.
Gender egalitarianism	4.70	0.35	Multigroup graded-response Rasch-based index (Warm, 1989) of 7-point Likert responses to five questions (See Appendix A; House et al., 2004). Min = 3.89; Max = 5.17; Reliability = .95.
Uncertainty avoidance	2.57	0.68	Multigroup graded-response Rasch-based index (Warm, 1989) of 7-point Likert responses to five questions (See Appendix A; House et al., 2004). Min = 1.39; Max = 3.84; Reliability = .96.
Clustering of students by parent rating of past literacy skills	0.15	0.09	Ratio of parent rating of past literacy skills variance across schools/country variance of parent rating of past literacy skills. Min = 0.06; Max = 0.36.
Family variables (entered at the student level)			
Family SES ^a	0.00	1.00	Index of SES: Father education; mother's education; father's occupation; mother's occupation; family's financial situation. Min = -2.79; Max = 3.06; Reliability = .94.
Home education resources ^a	0.00	1.00	Index of availability of educational resources at home: computer; study desk/table for student's own use; books of his/her own; access to a daily newspaper (choices: yes, no); number of books at home (choices: 0-10, 11-25, 26-100, 101-200, >200); no. of children's books at home (choices: 0-10, 11-25, 26-50, 51-100, >100). Min = -2.52; Max = 2.05; Reliability = .79.
Parent attitude toward reading ^a	0.00	1.00	Index of statements: "I read only if I have to"; "I like talking about books with other people"; "I like to spend my spare time reading"; "I read only if I need information"; "Reading is an important activity in my home." Response choices = <i>disagree a lot, disagree a little, agree a little, agree a lot</i> . Min = -2.97; Max = 1.67; Reliability = .82.
School variables (entered at the class level)			
Availability of school resources ^a	0.00	1.00	Inverted index of shortage of: Qualified teaching staff; teachers with a specialization in reading; instructional materials; supplies; school buildings and grounds; heating/cooling and lighting systems; instructional space; computers for instructional purposes; computer software for instructional purposes; library books; audio-visual resources. Response choices: <i>not at all, a little, some, a lot</i> . Min = -2.71; Max = 1.09; Reliability = .94.

(table continues)

Table 2 (Continued)

Variable	M	SD	Description
Class mean variables (entered at the class level)			
Class mean parent rating of past literacy skills ^a	0.00	0.48	Class mean of students' past achievement. Min = -2.28, Max = 2.02.
Class mean SES ^a	0.00	0.62	Class mean of parents' SES. Min = -2.79; Max = 2.20.
Class mean home education resources ^a	0.00	0.71	Class mean of home education resources. Min = -2.52; Max = 2.05.
Class mean parents attitude toward reading ^a	0.00	0.37	Class mean of parents' attitude toward reading. Min = -2.97; Max = 1.67; Reliability = .82.
Psychology variables (entered at the student level)			
Student reading attitude ^a	0.00	1.00	Index of statements: "I read only if I have to"; "I like talking about books with other people"; "I would be happy if someone gave me a book as a present"; "I think reading is boring"; "I enjoy reading." Response choices = <i>disagree a lot, disagree a little, agree a little, agree a lot</i> . Min = -2.66; Max = 0.98; Reliability = .86.
Student reading self-concept ^a	0.00	1.00	Index of statements: "Reading is very easy for me"; "I do not read as well as other students in my class"; "Reading aloud is very hard for me." Response choices = <i>disagree a lot, disagree a little, agree a little, agree a lot</i> . Min = -2.55; Max = 1.53; Reliability = .62.
Gender (entered at the student level)			
Girl	0.50		1 = girl; 0 = boy.

Note. Data are from Progress in International Reading Literacy Study (PIRLS), unless otherwise noted. GDP = gross domestic product; SES = socioeconomic status; Min = minimum; Max = maximum.

^a Indices were standardized to $M = 0$ and $SD = 1$.

data issues more effectively than deletion, mean substitution, or simple imputation (Peugh & Enders, 2004).

Variables

The outcome variable is reading achievement. Explanatory variables include parent rating of past literacy skills and explanatory variables at the country, family, school, classmate, and student levels.

Reading achievement. The PIRLS framework defines the two major aspects of students' reading literacy—reading purposes and comprehension processes (Martin et al., 2003). Reading for literary experience and reading to acquire and use information are the two major purposes that account for the majority of reading experiences of young children. Thus, the questions are divided equally so that 50% address each purpose. Readers make meaning of texts in many ways, depending not only on the purpose for reading but also on the difficulty of the text and the reader's prior knowledge. PIRLS looks at four processes of comprehension: focusing on and retrieving explicitly stated information (20% of the questions); making straightforward inferences (30%); interpreting and integrating ideas and information (30%); and examining and evaluating content, language, and textual elements (20%).

International experts from PIRLS countries defined reading achievement, built assessment frameworks, created test items and questionnaire items, forward- and backward-translated them, and pilot tested them to check their validity and reliability (for details and sample items, see Martin et al., 2003, and www.pirls.org). Students did not respond to all items on the entire test. Instead, they received subtests (overlapping subsets of all multiple-choice and open-ended questions) for wider coverage of reading skills while reducing student fatigue and learning during the test (balanced incomplete block test; Baker & Kim, 2004). A graded-response Rasch model of these subtests measured the difficulty of each test item to estimate each student's reading competence more precisely (Baker & Kim, 2004).

Parent rating of past literacy skills. This graded-response Rasch-based index (using the Warm, 1989, procedure) was created from a parent's or guardian's responses to multiple questions in order to reduce measurement error. The items used to create this index were "My child . . ." (a) recognizes most of the alphabet letters, (b) reads words, (c) reads sentences, (d) writes letters of the alphabet, and (e) writes words. The response choices were *not at all, not very well, moderately well, and very well*. Its reliability was 0.95. All reliabilities were measured with Cronbach's alpha.

Country. Country-level variables included economic conditions, cultural values, and distribution of students across schools. Economic growth was measured through gross domestic product per capita (GDP per capital; World Bank, 2002). Family income inequality was measured through GDP Gini index (the integral of the cumulative distribution function of a perfectly equal income society minus the integral of the cumulative distribution function of the actual society's income; World Bank, 2002). Scores can range from 0 (*perfect equality; everyone has equal income*) to 100 (*perfect inequality; one person has all the income, and everyone else's income is zero*). The Gini index is suitable for nonnormal distributions like household income (McKenzie, 2005).

Similar to the Organization for Economic Cooperation and Development's (OECD) Program for International Student Assess-

ment (OECD, 2010), a consortium of 150 researchers collected over 17,300 responses to a survey of cultural values from middle managers in finance, telecommunications, and food processing in 61 cultures (House et al., 2004). Cultural values differ mostly across countries, not within countries; indeed, cultural values are linked more strongly to one's nation than to religion, employer organization, or individual personality (Hofstede, Neuijen, Ohayv, & Sanders, 1990; Inglehart & Baker, 2000), so managers' cultural values serve as proxies for those of the entire nation. House et al. (2004) created all indices of cultural values from a polychoric correlation-based factor analyses of managers' responses to questions on a 7-point Likert scale.

Cultural values included power distance, in-group collectivism, gender egalitarianism, and uncertainty avoidance. Hierarchy, or *power distance*, is the degree to which people expect and agree that power should be shared unequally, based on five questions with a reliability of 0.88. (see questionnaire items for cultural values in Appendix A, House et al., 2004). *In-group collectivism* is the degree to which people value collective action and collective distribution of resources, based on four questions with a reliability of 0.95. Meanwhile, *gender egalitarianism* is the degree to which people minimize gender inequality, based on five questions with a reliability of 0.95. *Uncertainty avoidance* is the degree to which people rely on social norms, rules, and procedures to reduce the unpredictability of future events, based on five questions with a reliability of 0.96.

The distribution of students across schools within a country can differ. In some countries, high-achieving students attend one set of schools while low-achieving students attend a different set of schools, resulting in high clustering of students by past achievement. In other countries, students are mixed together in the same schools regardless of past achievement. Hence, *clustering of students by parent rating of past literacy skills* is the ratio of parent rating of past literacy skills variance across schools divided by the total parent rating of past literacy skills variance within a country.

Family. Family variables included three graded-response Rasch-based indices of parent responses to questionnaire items: SES, home educational resources, and parent attitude toward reading. *SES* was created from father's education, mother's education, father's occupation, mother's occupation, and responses to a question on family financial situation ("How well off do you think your family is financially?" with the response choices of *not at all well off*, *not very well off*, *average*, *somewhat well off*, and *very well off*). Occupation responses were recoded according to job status (according to Ganzeboom, 1992). Its reliability was 0.94.

Home educational resources was created from the availability of the following educational resources at home: computer, study desk or table for the student's own use, books of his or her own, access to a daily newspaper (choices: yes or no), number of books at home (choices: 0–10, 11–25, 26–100, 101–200, and > 200); number of children's books at home (choices: 0–10, 11–25, 26–50, 51–100, and > 100). Its reliability was 0.79.

Parent attitude toward reading was created from responses to the following questions: "I read only if I have to"; "I like talking about books with other people"; "I like to spend my spare time reading"; "I read only if I need information"; and "Reading is an important activity in my home." The possible response choices were *disagree a lot*, *disagree a little*, *agree a little*, and *agree a lot*. Its reliability was 0.82.

Availability of school resources was an inverted graded-response Rasch-based index of whether the principal perceived a shortage of the following: qualified teaching staff; teachers with a specialization in reading; instructional materials, supplies; school buildings and grounds; heating/cooling and lighting systems; instructional space; computers for instructional purposes; computer software for instructional purposes; library books; and audio-visual resources. Response choices were *not at all*, *a little*, *some*, and *a lot*. Its reliability was 0.94.

Classmates. To test whether characteristics of classmates are linked to a student's reading achievement, we included both the characteristic of a student (e.g., SES) and the mean of this characteristic for all students in the same class (class mean SES) into a regression. By controlling for SES, the regression coefficient of class mean SES indicates the relationship between classmates' SES and a student's reading achievement. Similarly, we computed class mean parent rating of past literacy skills, class mean home educational resources, and class mean attitude toward reading.

Psychology and gender. Student variables included student reading attitude, student reading self-concept, and girl. *Student reading attitude* was a graded-response Rasch-based index of student responses to the following: "I read only if I have to"; "I like talking about books with other people"; "I would be happy if someone gave me a book as a present"; "I think reading is boring"; and "I enjoy reading." Response choices were *disagree a lot*, *disagree a little*, *agree a little*, and *agree a lot*. Its reliability was 0.86. *Student reading self-concept* was a graded-response Rasch-based index of student responses to the following: "Reading is very easy for me"; "I do not read as well as other students in my class"; and "Reading aloud is very hard for me." Response choices were *disagree a lot*, *disagree a little*, *agree a little*, and *agree a lot*. As its reliability was 0.62, results involving this variable require cautious interpretation. Last, *girl* has a value of 1 for girls and a value of 0 for boys. See Table 2 for overall summary statistics and Appendix Tables B1, B2, and B3 for correlation-variance-covariance matrices; country means of (a) numbers of classes, (b) numbers of students, (c) percentage of missing data; and (d) key variables.

Analysis. A multilevel logit analysis of plausible values yields more precise standard errors than does ordinary least squares (Goldstein, 1995; Rust & Rao, 1996). The simple variance components multilevel model tests if the variance at each level is significant.

$$\text{Reading}_{ijk} = \beta_{000} + e_{ijk} + f_{0jk} + g_{00k} \quad (1)$$

The outcome variable Reading_{ijk} of student i in school j in country k has a grand mean intercept β_{000} , with student-, school-, and country-level residuals (e_{ijk} , f_{0jk} , and g_{00k} , respectively). Explanatory variables were entered in sequential sets to estimate the variance explained by each set (Kennedy, 2008). The index of a student's early literacy skills at first grade (see Table 2 for details) reflects the cognitive component of reading and past reading skills and is entered first. Next, we considered ecological variables, namely, country, family, classmate, and school characteristics. Country variables might affect family variables. As families might choose their children's schools, family variables might affect classmate and school variables. All of these variables, broadly defined as the "ecology" in which reading occurs, might affect

students' psychological variables. Hence, we entered the variables as follows: parent rating of past literacy skills, ecological (country, family, classmate, school), and psychological. All continuous variables were centered on their country mean.

$$\begin{aligned}
 \text{Reading}_{ijk} = & \beta + e_{ijk} + f_{0jk} + g_{00k} \\
 & + \beta_{1jk} \text{Prior_literacy_skills} \\
 & + \beta_c \text{Country} + \beta_{fjk} \text{Family}_{ijk} + \beta_{mk} \text{Classmate}_{jk} \\
 & + \beta_{sk} \text{School}_{jk} + \beta_{pj} \text{Psychological}_{ijk} + \beta_{2jk} \text{Girl}_{ijk}
 \end{aligned} \quad (2)$$

First, we entered each student's index of parent rating of past literacy skills when they began first grade (*Prior_literacy_skill*, see Table 2). Then, we tested whether sets of predictors were significant with a nested hypothesis test (chi-square log likelihood; Kennedy, 2008). Nonsignificant variables were removed. Then, we applied this procedure for *Prior_literacy_skill* to the country variables, log GDP per capita, Gini index, clustering of students by past achievement, power distance, in-group collectivism, gender egalitarianism, and uncertainty avoidance (**Country**). Next, we applied this procedure to family variables: socioeconomic status (SES), home education resources, and parent attitude toward reading (**Family**).

To test whether country economic characteristics or cultural values moderate these links, we applied a random effects model (Goldstein, 1995) to determine if the regression coefficients ($\beta_{fjk} = \beta_{v00} + f_{fjk} + g_{f0k}$) differed across countries ($g_{f0k} \neq 0$) or correlated with **Country** variables.

Then, we applied the procedure for **Family** to classmates' family variables: class mean SES, class mean home education resources, class mean attitude toward reading, and class mean parent rating of past literacy skills (**Classmate**). This part of the specification tests whether classmates' family SES, home literacy resources, attitude toward reading, or their past reading achievement affect a student's reading achievement. As noted previously, the random effects model tests if country characteristics moderate links between classmate characteristics and student reading achievement.

Next, we applied this procedure to the school variable: availability of school resources (**School**). Then, we applied this procedure to psychological variables: attitude toward reading, reading self-concept (**Psychological**). We also tested gender (**Girl**).

Furthermore, we tested whether the relation between classmate characteristics and reading achievement differed across students by applying a random effects model (Goldstein, 1995). We tested if the regression coefficients ($\beta_{fjk} = \beta_{v00} + f_{fjk} + g_{f0k}$; $\beta_{pj} = \beta_{v00} + f_{pj} + g_{f0k}$) differed across classes ($f_{fjk} \neq 0$; $f_{pj} \neq 0$) or correlated with **Classmate** variables.

To test for multilevel, mediation effects by classmate variables, we used the multilevel M test, which corrects for potential non-normal distributions and tests the significance of a confidence interval based on a critical *z* ratio determined across multiple data simulations (MacKinnon, Lockwood, & Williams, 2004). This test reduces false positives (MacKinnon et al., 2004) and has more power than other methods to detect small mediation effects (Pituch, Stapleton, & Kang, 2006).

We report how a 10% increase in each continuous variable above its mean is linked to reading achievement (result = $b * SD * [10\% / 34\%]$; $1 SD \approx 34\%$). As percentage increase is not linearly related to standard deviation, scaling is not warranted.

We used an alpha level of .05. To minimize false positives, we controlled for the false discovery rate with the two-stage linear step-up procedure, which outperformed 13 other methods in computer simulations (Benjamini, Krieger, & Yekutieli, 2006). The small sample of countries ($N = 33$) limits identification of non-significant country-level results (Konstantopoulos, 2008; see Table 3 for details). As our random effects model portion of the multilevel analysis produces estimates of the effects for each country, we tested whether the results differed across the 33 subsamples of each country. We also analyzed residuals for influential outliers. The analyses were completed with item response theory (IRT) command language (Hanson, 2002), Mln (Rasbash & Woodhouse, 1995), and LISREL (Jöreskog & Sörbom, 2004).

Results

Explanatory Model

Parent rating of past literacy skills, country, family, classmate, school, and psychological variables accounted for differences in students' reading scores (see Table 4). Reading scores differed across countries (32%), across schools (20%), and across students (47% of the variance). All results discussed in this section describe first entry into the regression, with all previously included variables controlled. Ancillary regressions and statistical tests are available upon request.

Parent rating of past literacy skills. As expected, students with higher parent ratings of past literacy skills had higher reading achievement scores. Parent rating of past literacy skills accounted for 4% of the variance in students' reading achievement.

Country. Countries' economic characteristics and degree of clustering of students, but not its cultural variables, were linked to a student's reading achievement. Students in countries with greater economic growth (higher GDP per capita) had higher reading scores (*log GDP per capita* accounted for more variance than *GDP per capita*). Economic growth showed the strongest link with reading achievement, with a beta of 0.27 in the final model (see Table 4, Model 4, row 8). In countries with greater family income inequality (higher Gini), students had lower reading scores. Also, in countries with greater clustering of students across schools by parent rating of past literacy skills (tracking or banding), students had lower reading scores. Cultural values were not directly related to reading scores. These country variables accounted for 9% of the

Table 3
Statistical Power at Each Level of Analysis for Each Effect Size

Level-variable	Effect size			
	0.1	0.2	0.3	0.4
3-Country	0.09	0.20	0.39	0.61
2-Class	0.65	1.00	1.00	1.00
1-Student	0.71	1.00	1.00	1.00

Note. Sample = 141,019 fourth-grade students from 5,279 schools in 33 countries.

Table 4

Summary of Four Multilevel Regression Models Predicting Students' Reading Scores: Unstandardized Regression Coefficients, (Standard Errors), and Standardized Regression Coefficients

Explanatory variable	Regressions predicting reading achievement			
	Model 1	Model 2	Model 3	Model 4
Parent rating of past literacy skills	14.83 (0.15)***	14.82 (0.15)***	8.82 (0.15)***	8.97 (0.15)***
Log GDP per capita		43.26 (4.62)***	42.21 (3.72)***	38.98 (3.65)***
Gini index		-2.78 (0.37)***	-1.59 (0.30)***	-1.61 (0.29)***
Clustering of students across schools by parent rating of past literacy skills	-107.8 (29.76)***	-92.23 (29.20)**	-39.56 (28.69)	
Power distance	0.10 (0.14)	0.11 (0.12)	0.09 (0.11)	
In-group collectivism	-0.12 (0.14)	-0.05 (0.12)	-0.11 (0.11)	
Gender egalitarianism	0.11 (0.08)	0.13 (0.07)	0.06 (0.07)	
Uncertainty avoidance	0.02 (0.11)	0.02 (0.09)	0.01 (0.09)	
Socioeconomic status		8.54 (0.17)***	7.97 (0.18)***	
Home education resources		10.40 (0.20)***	10.01 (0.20)***	
Parent attitude toward reading		3.42 (0.14)***	3.22 (0.14)***	
Class mean socioeconomic status		11.52 (0.94)***	10.83 (0.93)***	
Class mean home education resources		19.67 (1.06)***	20.28 (1.05)***	
Class mean attitude towards reading		9.19 (0.81)***	8.50 (0.81)***	
Class mean parent rating of past literacy skills		5.37 (0.91)***	6.62 (0.92)***	
Availability of school resources		1.30 (0.51)*	1.16 (0.49)*	
Students' attitude towards reading		9.30 (0.14)***	9.47 (0.15)***	
Students' reading self-concept		16.49 (0.14)***	16.35 (0.14)***	
Girl		8.71 (0.28)***	8.65 (0.28)***	
Log GDP per capita * Class mean parent rating of past literacy skills				6.54 (1.91)**
In-group collectivism * Class mean parent rating of past literacy skills				0.37 (0.06)***
Uncertainty avoidance * Class mean parent rating of past literacy skills				-0.44 (0.05)***
SES * Class mean SES				1.85 (0.24)***
Home education resources * Class mean home education resources				3.57 (0.26)***
Attitude towards reading * Class mean attitude towards reading				2.48 (0.38)***
Parent rating of past literacy skill * Class mean parent rating of past literacy skills				5.31 (0.31)***
Variance at each level				
Country (32%)	0.02	0.07	0.42	0.42
School (20%)	0.06	0.06	0.43	0.46
Student (47%)	0.04	0.04	0.20	0.20
Total variance explained	0.04	0.05	0.31	0.32

Note. Each regression included a constant term. GDP = gross domestic product; SES = socioeconomic status.

* $p < .05$. ** $p < .01$. *** $p < .001$.

reading achievement differences between countries and 1% of the total variance in students' reading achievement.

Family. Students with higher family SES, more educational resources at home, or better parent attitudes toward reading scored higher in reading, consistent with past research. Family characteristics accounted for 15% of the variance in reading scores.

Classmates. When classmates had higher family SES, more home education resources, better parent reading attitudes, better reading attitudes, or greater parent rating of past literacy skills, a student scored higher in reading. These results support the view that classmate family factors contribute to a student's reading achievement. Classmates' home education resources showed the third strongest link to a student's reading achievement ($\beta = .14$). Controlling for classmate family factors, classmate attitudes and parent rating of past literacy skills were still significantly linked to student reading achievement, showing that classmate family factors do not fully explain the relation between classmates and a student's reading achievement. Notably, the regression coefficient of class mean attitude toward reading was larger than that of class mean parent rating of past literacy skills. Classmate characteristics accounted for 6% of the variance in reading scores.

School. Furthermore, students in schools with more resources had higher reading scores. School resources accounted for 2% of the variance in reading scores.

Psychology and gender. Students with better reading attitudes or higher reading self-concept (second highest $\beta = .17$) had higher reading scores, accounting for 2% of the variance in reading scores. Girls outscored boys in reading on average, accounting for 1% of the variance in reading scores.

Differences across countries. The link between reading achievement and class mean past achievement varied across countries' economic status and cultural contexts. In countries that were wealthier, more collectivist, or less uncertainty avoidance, the link between classmate parent rating of past literacy skills and a student's reading score was larger.

Differences across students. The links between classmate characteristics (SES, home educational resources, attitude toward reading, and parent rating of past literacy skills) and a student's reading achievement also varied across students. With respect to reading achievement score, a higher SES student benefits more than lower SES students from higher SES classmates. Likewise, when classmates have greater home education resources, students with more such resources benefit more than

students with fewer such resources. In schools where classmates have better attitudes toward reading, students who have better attitudes toward reading benefit more than those with weaker attitudes toward reading. Last, students with stronger parent rating of past literacy skills benefit more than students with weaker such skills from classmates with greater parent rating of past literacy skills. Note that these interactions do not nullify the large regression coefficient of class mean SES. Despite the differences in benefits, low-SES students with high-SES classmates still have substantially higher reading test scores than other students. In short, when classmates have more of a resource, attitude, or skill, students with more of it benefit more than students with less of it. Thus, these results are specific to each dimension (students with greater parent rating of past literacy skills did *not* benefit more from classmates with better attitudes toward reading).

These relationships across countries and across students accounted for 1% of the variance in reading scores. Otherwise, these results were consistent across all 33 countries, and there was no significant mediation. Examination of the residuals did not show influential outliers.

Discussion

While students with higher achieving classmates show higher academic achievement in many countries with different education policies (Kang, 2007; Zimmer & Toma, 2000), researchers have not explicated whether classmate characteristics are related to a student's learning nor have they tested whether these links differ across countries. The present study extends this line of research by showing that classmate characteristics (parent rating of past literacy skills, attitudes toward reading) and classmates' parents' characteristics and resources (parent SES, material educational resources at home, parents' attitudes toward reading) were both associated with greater student reading achievement in 33 countries, controlling for family characteristics, which were also significantly and substantially related to reading achievement (consistent with past studies; e.g., Chiu & McBride-Chang, 2006). Furthermore, classmates' parent rating of past literacy skills showed stronger links to student reading achievement in countries that are richer, more collectivist, or more tolerant of uncertainty. These findings underscore the importance of multiple levels of environmental contexts and their interactions to provide a more comprehensive account of children's academic achievement. We discuss each of these findings.

Classmates' attitudes toward reading and parent rating of past literacy skills were both linked to a student's reading achievement. The standardized regression coefficient of classmates' attitudes toward reading is larger than that of classmate parent rating of past literacy skills, suggesting that classmate attitude has a stronger link with a student's reading achievement. Possibly, classmate achievement might be more likely than classmate attitude to trigger a negative social comparison, which could reduce a student's reading self-concept and, consequently, his or her reading achievement (Chiu & Klassen, 2009). Future studies can further examine this issue.

Classmates' parent and home characteristics were also related to a student's reading achievement. Consistent with past research, students whose classmates had higher family SES had higher

reading performance (Caldas & Bankston, 1997). Controlling for classmate parent SES, students whose classmates had more educational resources at home had higher reading achievement. Classmate home educational resources had the third largest standardized regression coefficient (behind log GDP per capita and reading self-concept). Both classmate SES and classmate home educational resources had much larger standardized regression coefficients than classmate attitude or classmate parent rating of past literacy skills, suggesting that classmate family resources have stronger links to a student's learning than do attitude or literacy skills. These findings highlight the prominence of classmates' family environments (Bradley & Corwyn, 2002). Such family environments might help both their children's reading achievement directly and their classmates' reading achievements indirectly; thus, their benefits are twofold.

However, classmates did not benefit each student equally. When classmates had more of a resource, attitude, or skill, students with more of it had higher reading achievement than students with less of it. These dimension-specific results are consistent with the dimension-specific mechanisms of selective trading, norm establishment/maintenance, and skill practice. Selective trading of educational resources might have helped students from high-SES families or with more home educational resources benefit from similarly advantaged classmates (Mankiw, 2011). When classmates have positive attitudes toward reading, students with positive attitudes toward reading might engage in normative reading-related practices more often and learn more, compared with students with negative attitudes toward reading (Chiu & McBride-Chang, 2006). Moreover, classmates with greater parent rating of past literacy skills might engage and benefit students with stronger such skills more than those with weaker such skills.

This study also extends past international research on classmates' relations to reading achievement by examining how these relations differ across countries. Specifically, the present study indicated that the links between classmate parent rating of past literacy skills and a student's current reading achievement was stronger in countries that were richer, more collectivist, or more tolerant of uncertainty. The stronger classmate link in richer countries is consistent with the view that classmates in richer countries have more resources that they can use to influence a student's reading achievement (Chiu & Chow, 2010). Meanwhile, the stronger classmate link in more collectivist countries fits the view that classmates in more collectivist cultures are more likely to pay attention to one another, interact with one another, help one another, and hence influence one another (Wade-Benzoni et al., 2002). Also, the weaker classmate link in countries with greater uncertainty avoidance is consistent with the view that students interact with their classmates less often and learn less from them in countries with greater uncertainty avoidance.

Implications

If future studies replicate these results, they would have several implications for research and policy: (a) multidimensional classmate links to student learning, (b) possible classmate interventions, (c) potential for harmful economic segregation of students, (d) unequal classmate benefits, (e) dimension-specific advantages, and (f) country differences. First, the multidimensional characteristics of classmates and their families are necessary components of a complete

theoretical model of classmates and student learning. A student does not simply benefit from classmate resources in a general, generic way. Instead, each specific classmate resource has a specific benefit. Furthermore the relationships between each classmate characteristic and a student's reading achievement differed. Indeed, classmate family characteristics (classmate family SES, home educational resources) had stronger links to student reading achievement than classmate attitudes toward reading did, which, in turn, had a stronger link to it than classmate parent rating of past literacy skills did.

Second, the impact of classmates suggests that interventions for low-achieving students might consider including classmates. Beyond serving as sources of information, classmates might help a student through motivating a better attitude toward reading, greater study time, or further perseverance. Classmates might also help create and maintain supportive norms of attitude, behavior, and achievement. Future studies can test such interventions.

Third, the substantial standardized regression coefficients of classmate family SES and home educational resources (especially compared to classmate attitude or parent rating of past literacy skills) are consistent with the danger of harmful economic segregation (Chiu & Khoo, 2005). If rich students attend separate schools away from poor students, poor students can lose access to rich students' family resources and learn less (Chiu, *in press*).

Fourth, even if students are economically integrated into the same schools, they do not share resources equally (Ryan, 2001). High-SES classmates benefit high-SES students more than low-SES students. Likewise, classmates with more home educational resources benefit students with more home educational resources more than those with fewer such resources. These results are consistent with: (a) high-SES students have more resources or skills to attract other high-SES classmates (Chiu & Chow, 2010) or (b) dimension-specific advantages through selective trading of resources (Mankiw, 2011).

Fifth, classmate attitudes and skills show specific benefits. These results suggest that we match classmates with, say, specific skills to students who need those skills rather than simply letting a student get help from any classmate who might have the desired skills. Furthermore, classmates with better attitudes toward reading benefit students with better attitudes toward reading more than those with poorer attitudes, possibly through classmate norms that encourage related learning behaviors (Chiu & McBride-Chang, 2006). Meanwhile, classmates' parent rating of past literacy skills benefit students with greater parent rating of past literacy skills more than those with weaker such skills, possibly through trading information or shared practice (Mankiw, 2011). Future studies can test these possible mechanisms.

Sixth, the standardized regression coefficients of classmate characteristics differ across countries with respect to their economies and cultural values. Therefore, educational interventions and policies that rely on peer interaction and cooperation at school might have greater impact in countries that are richer, are more collectivist, or avoid uncertainty less. These findings underscore the view that a full account of classmate influence must capture characteristics of each country's economy and cultural contexts.

Still, most of the relations between classmate characteristics and reading achievement were consistent and robust across countries. (Classmate literacy skills is the notable exception.) Thus, these classmate relations with reading achievement remain candidates

for universality and show the importance of classmates when seeking to understand antecedents of reading achievement. Overall, the findings of this study showed academic achievement's links with family, classmate, and country characteristics, underscoring the importance of studying various ecological systems as conceptualized in Bronfenbrenner's (2005) theory.

Limitations and Future Studies

This study had several limitations. First, this study focused on students' reading achievement, in which students learn different scripts across countries. Future studies can include different aspects of academic performance, such as mathematics and history, to provide a more comprehensive picture of classmate influence. Second, some measures in this pre-existing data do not perfectly capture the constructs of interest. For example, the cultural values are similar among citizens within a country (Inglehart & Baker, 2000), but some students, especially from cultural minorities, might have substantially different cultural values. For example, a Chinese American student living in the United States might have much more collectivist values than most individualistic U.S. students. As this study does not account for these possibilities, the statistical power of this study is lower, possibly yielding some nonsignificant statistical results that in reality should be significant. Thus, future studies can also collect individual students' cultural values. Also, some of the differences in reading achievement between countries might stem from their different orthographies. Specifically, the difficulty levels of reading vary across orthographies, with the transparent scripts easier to learn than the opaque ones. Also, the timing of when children formally start learning to read differs across countries. Furthermore, future studies can use behavioral measures rather than surveys. Last, this correlational study does not warrant causal interpretations, which future studies can help address with longitudinal designs.

Conclusion

This study explicated the relationship between classmate reading achievement and student reading achievement, suggesting possible mechanisms and showing when these differed across countries. Classmates' family SES and educational resources at home were more strongly linked to student reading achievement than were classmates' attitudes toward reading or their parent rating of past literacy skills, controlling for country, family, school, and student characteristics. However, these classmate links to reading achievement differed across students. High-SES classmates benefited high-SES students more than low-SES students. These dimension-specific advantages also applied to classmates' home educational resources, attitudes toward reading, and parent rating of past literacy skills.

Furthermore, the links between classmate past reading achievement and a student's current reading achievement were stronger in countries that were richer, were more collectivist, or had less uncertainty avoidance. Other classmate characteristics' links to a student's past achievement did not differ significantly across countries, so they remain candidates for universal relations across all countries. These findings underscore the view that a full account of a student's reading achievement must capture the ecological relationships in the family microsystem, classmate microsystem, classmate family mesosystem, and the country macrosystem.

References

Baker, D. P., Goesling, B., & Letendre, G. K. (2002). Socioeconomic status, school quality, and national economic development. *Comparative Education Review*, 46, 291–312. doi:10.1086/341159

Baker, F. B., & Kim, S.-H. (2004). *Item response theory*. Boca Raton, FL: CRC Press.

Baker, L., & Wigfield, A. (1999). Dimensions of children's motivation for reading and their relations to reading activity and reading achievement. *Reading Research Quarterly*, 34, 452–477. doi:10.1598/RRQ.34.4.4

Benjamini, Y., Krieger, A. M., & Yekutieli, D. (2006). Adaptive linear step-up procedures that control the false discovery rate. *Biometrika*, 93, 491–507. doi:10.1093/biomet/93.3.491

Bradley, R. H., & Corwyn, R. F. (2002). Socioeconomic status and child development. *Annual Review of Psychology*, 53, 371–399. doi:10.1146/annurev.psych.53.100901.135233

Bronfenbrenner, U. (2005). *Making human beings human: Bioecological perspectives on human development*. Thousand Oaks, CA: Sage.

Brown, G., Micklewright, J., Schnepf, S. V., & Waldmann, R. (2007). Cross-national surveys of learning achievement. *Journal of the Royal Statistical Society, Series A: Statistics in Society*, 170, 623–646. doi:10.1111/j.1467-985X.2006.00439.x

Caldas, S. J., & Bankston III, C. (1997). Effect of school population socioeconomic status on individual academic achievement. *Journal of Educational Research*, 90, 269–277. doi:10.1080/00220671.1997.10544583

Chiu, M. M. (2007). Families, economies, cultures, and science achievement in 41 countries. *Journal of Family Psychology*, 21, 510–519. doi:10.1037/0893-3200.21.3.510

Chiu, M. M. (2008). Flowing toward correct contributions during groups' mathematics problem solving: A statistical discourse analysis. *Journal of the Learning Sciences*, 17, 415–463. doi:10.1080/10508400802224830

Chiu, M. M. (2010). Inequality, family, school, and mathematics achievement. *Social Forces*, 88, 1645–1676. doi:10.1353/sof.2010.0019

Chiu, M. M. (2013). Family structure and education. In J. Ainsworth & G. J. Golson (Eds.), *Sociology of education* (pp. 271–275). Thousand Oaks, CA: Sage.

Chiu, M. M. (in press). Family inequality, school inequalities, economic segregation and mathematics achievement: SES and ability segments of fifteen-year-olds in 65 countries. *Teacher's College Record*.

Chiu, M. M., & Chen, G. (in press-a). Collectivism. In L. Ganong, M. Coleman & G. J. Golson (Eds.), *The social history of the American family: An encyclopedia*. Thousand Oaks, CA: Sage.

Chiu, M. M., & Chen, G. (in press-b). Individualism. In L. Ganong, M. Coleman & G. J. Golson (Eds.), *The social history of the American family: An encyclopedia*. Thousand Oaks, CA: Sage.

Chiu, M. M., & Chow, B. W. Y. (2010). Culture, motivation, and reading achievement. *Learning and Individual Differences*, 20, 579–592. doi:10.1016/j.lindif.2010.03.007

Chiu, M. M., Chow, B. W.-Y., & McBride-Chang, C. (2007). Universals and specifics in learning strategies. *Learning and Individual Differences*, 17, 344–365. doi:10.1016/j.lindif.2007.03.007

Chiu, M. M., & Ho, S. C. (2006). Family effects on student achievement in Hong Kong. *Asia Pacific Journal of Education*, 26, 21–35. doi:10.1080/02188790600607846

Chiu, M. M., & Khoo, L. (2003). Rudeness and status effects during group problem solving: Do they bias evaluations and reduce the likelihood of correct solutions? *Journal of Educational Psychology*, 95, 506–523. doi:10.1037/0022-0663.95.3.506

Chiu, M. M., & Khoo, L. (2005). Effects of resources, inequality, and privilege bias on achievement. *American Educational Research Journal*, 42, 575–603. doi:10.3102/00028312042004575

Chiu, M. M., & Klassen, R. M. (2009). Calibration of reading self-concept and reading achievement among 15-year-olds. *Learning and Individual Differences*, 19, 372–386. doi:10.1016/j.lindif.2008.10.004

Chiu, M. M., & McBride-Chang, C. (2006). Gender, context, and reading: A comparison of students in 43 countries. *Scientific Studies of Reading*, 10, 331–362. doi:10.1207/s1532799xssr1004_1

Chiu, M. M., & Zeng, X. (2008). Family and motivation effects on mathematics achievement. *Learning and Instruction*, 18, 321–336. doi:10.1016/j.learninstruc.2007.06.003

Coleman, J. S. (1994). Family, school, and social capital. In T. Husen & T. N. Postlethwaite (Eds.), *International encyclopedia of education* (2nd ed., pp. 2272–2274). Oxford, England: Pergamon Press.

Crosnoe, R. (2004). Social capital and the interplay of families and schools. *Journal of Marriage and Family*, 66, 2, 267–280. doi:10.1111/j.1741-3737.2004.00019.x

Davalos, D. B., Chavez, E. L., & Guardiola, R. J. (2005). Effects of perceived parental school support and family communication on delinquent behaviors in Latinos and White non-Latinos. *Cultural Diversity and Ethnic Minority Psychology*, 11, 57–68. doi:10.1037/1099-9809.11.1.57

Edmunds, K. M., & Bauserman, K. L. (2006). What teachers can learn about reading motivation through conversations with children. *The Reading Teacher*, 59, 414–424. doi:10.1598/RT.59.5.1

Ganzeboom, H. B. G. (1992). A standard international socio-economic index of occupational status. *Social Science Research*, 21, 1–56. doi:10.1016/0049-089X(92)90017-B

Goldstein, H. (1995). *Multilevel statistical models*. Sydney, NSW, Australia: Arnold.

Guthrie, J. T., Klauda, S. L., & Morrison, D. A. (2012). Motivation, achievement, and classroom contexts for information book reading. In J. T. Guthrie, A. Wigfield, & S. L. Klauda (Eds.), *Adolescents' engagement in academic literacy* (pp. 1–51). College Park: University of Maryland.

Hanson, B. A. (2002). *IRT command language (ICL)* [Computer software]. Retrieved from <http://www.b-a-h.com/software/irt/icl/>

Hofstede, G. (2003). *Culture's consequences*. Thousand Oaks, CA: Sage.

Hofstede, G., Neuijen, B., Ohayv, D. D., & Sanders, G. (1990). Measuring organizational cultures. *Administrative Science Quarterly*, 35, 286–316.

House, R. J., Hanges, P. J., Javidan, M., Dorfman, P. W., & Gupta, V. (2004). *Culture, leadership, and organizations*. Thousand Oaks, CA: Sage.

Inglehart, R., & Baker, W. E. (2000). Modernization, cultural change, and the persistence of traditional values. *American Sociological Review*, 65, 19–51. doi:10.2307/2657288

Johnson, D. W., & Johnson, R. (1999). *Learning together and alone: Cooperative, competitive, and individualistic learning* (5th ed.). Boston, MA: Allyn & Bacon.

Jöreskog, K., & Sörbom, D. (2004). *LISREL Version 8.7* [Computer software]. New York, NY: Scientific Software.

Kang, C. (2007). Academic interactions among classroom peers: A cross-country comparison using TIMSS. *Applied Economics*, 39, 1531–1544. doi:10.1080/00036840600606328

Kennedy, P. (2008). *A guide to econometrics*. Cambridge, England: Blackwell.

Kindermann, T. A. (2007). Effects of naturally existing peer groups on changes in academic engagement in a cohort of sixth graders. *Child Development*, 78, 1186–1203. doi:10.1111/j.1467-8624.2007.01060.x

Konstantopoulos, S. (2008). The power of the test in three-level cluster randomized designs. *Journal of Research on Educational Effectiveness*, 1, 66–88. doi:10.1080/19345740701692522

LeTendre, G. K., Hofer, B. K., & Shimizu, H. (2003). What is tracking? *American Educational Research Journal*, 40, 43–89. doi:10.3102/00028312040001043

Lewis, R. (2001). Classroom discipline and student responsibility: The students' view. *Teaching and Teacher Education*, 17, 307–319. doi: 10.1016/S0742-051X(00)00059-7

Loera, G., Rueda, R., & Nakamoto, J. (2011). The association between parental involvement in reading and schooling and children's reading engagement in Latino families. *Literacy Research and Instruction*, 50, 133–155. doi:10.1080/19388071003731554

Ma, X., & Willms, J. D. (2004). School disciplinary climate: Characteristics and effects on eighth grade achievement. *Alberta Journal of Educational Research*, 50, 169–188.

MacKinnon, D. P., Lockwood, C. M., & Williams, J. (2004). Confidence limits for the indirect effect: Distribution of the product and resampling methods. *Multivariate Behavioral Research*, 39, 99–128. doi:10.1207/s15327906mbr3901_4

Mankiw, N. G. (2011). *Principles of economics* (6th ed.). Cincinnati, OH: South-Western College.

Martin, M. O., Mullis, I. V. S., & Kennedy, A. M. (2003). *PIRLS 2001 technical report*. Chestnut Hill, MA: Boston College.

May, H. (2006). A multilevel Bayesian item response theory method for scaling socioeconomic status in international studies of education. *Journal of Educational and Behavioral Statistics*, 31, 63–79. doi:10.3102/10769986031001063

McKenzie, D. (2005). Measuring inequality with asset indicators. *Journal of Population Economics*, 18, 229–260. doi:10.1007/s00148-005-0224-7

McPherson, M., Smith-Lovin, L., & Cook, J. M. (2001). Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 27, 415–444. doi:10.1146/annurev.soc.27.1.415

Murphy, T. (1994). Tests: Learning through negotiated interaction. *TESOL Journal*, 4, 12.

Opdenakker, M. C., & Van Damme, J. (2001). Relationship between school composition and characteristics of school process and their effect on mathematic achievement. *British Educational Research Journal*, 27, 407–432. doi:10.1080/01411920120071434

Organization for Economic Cooperation and Development. (2010). *PISA 2009 technical report*. Paris, France: Author.

Orr, A. J. (2003). Black–White differences in achievement: The importance of wealth. *Sociology of Education*, 76, 281–304. doi:10.2307/1519867

Pan, B. A., Perlmann, R. Y., & Snow, C. E. (2000). Food for thought. In L. Menn & N. B. Ratner (Eds.), *Methods for studying language production* (pp. 205–224). Mahwah, NJ: Erlbaum.

Peugh, J. L., & Enders, C. K. (2004). Missing data in educational research. *Review of Educational Research*, 74, 525–556. doi:10.3102/00346543074004525

Pituch, K. A., Stapleton, L. M., & Kang, J. Y. (2006). A comparison of single sample and bootstrap methods to assess mediation in cluster randomized trials. *Multivariate Behavioral Research*, 41, 367–400.

Pong, S.-L. (1997). Family structure, school context, and eighth-grade math and reading achievement. *Journal of Marriage and the Family*, 59, 3, 734–746. doi:10.2307/353957

Pong, S.-L. (1998). The school compositional effect of single-parenthood on 10th-grade achievement. *Sociology of Education*, 71, 23–43. doi:10.2307/2673220

Rasbash, J., & Woodhouse, G. (1995). *MLn command reference*. London, England: University of London Institute of Education, Multilevel Models Project.

Ream, R. (2003). Counterfeit social capital and Mexican-American under-achievement. *Educational Evaluations and Policy Analysis*, 25, 237–262. doi: 10.3102/01623737025003237

Rossi, P. E., Gilula, Z., & Allenby, G. M. (2001). Overcoming scale usage heterogeneity. *Journal of the American Statistical Association*, 96, 20–31. doi:10.1198/016214501750332668

Rust, K. F., & Rao, J. N. K. (1996). Variance estimation for complex estimators in sample surveys. *Statistics in Medical Research*, 5, 381–397.

Ryan, A. M. (2001). The peer group as a context for the development of young adolescent motivation and achievement. *Child Development*, 72, 4, 1135–1150. doi:10.1111/1467-8624.00338

Sallee, M. W., & Tierney, W. G. (2007). The influence of peer groups on academic success. *College and University*, 82, 7–14.

Schulz, W. (2003). *Validating questionnaire constructs in international studies*. Camberwell, VIC, Australia: Australian Council for Educational Research.

Skibbe, L. E., Phillips, B. M., Day, S. L., Brophy-Herb, H. E., & Connor, C. M. (2012). Children's early literacy growth in relation to classmates' self-regulation. *Journal of Educational Psychology*, 104, 541–553. doi:10.1037/a0029153

Wade-Benzoni, K. A., Okumura, T., Brett, J. M., Moore, D. A., Tenbrunsel, A. E., & Baserman, M. H. (2002). Cognitions and behavior in asymmetric social dilemmas: A comparison of two cultures. *Journal of Applied Psychology*, 87, 87–95. doi:10.1037/0021-9010.87.1.87

Walker, S. O., Petrill, S. A., & Plomin, R. (2005). A genetically sensitive investigation of the effects of school environment and socio-economic status on academic achievement at seven-year-olds. *Educational Psychology*, 25, 55–73. doi:10.1080/0144341042000294895

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, 54, 427–450. doi:10.1007/BF02294627

Willms, J. D. (1999). Quality and inequality in children's literacy: The effects of families, schools, and communities. In D. P. Keating & C. Hertzman (Eds.), *Developmental health and the wealth of nations* (pp. 72–93). New York, NY: Guilford Press.

World Bank. (2002). *The world development report 2001*. New York, NY: Oxford University Press.

Zimmer, R. W., & Toma, E. F. (2000). Peer effects in private and public schools across countries. *Journal of Policy Analysis and Management*, 19, 75–92. doi:10.1002/(SICI)1520-6688(200024)19:1<46::AID-PAM4>3.0.CO;2-Z

(Appendices follow)

Appendix A

GLOBE Cultural Value Questions

All questions below are on a 7-point Likert scale.

Power Distance

3-5. I believe that a person's influence in this society should be based primarily on:

(1) one's ability and contribution to the society . . . (7) the authority of one's position.

3-13. I believe that followers should:

(1) obey their leader without question . . . (7) question their leader when in disagreement.

3-28. I believe that people in positions of power should try to:

(1) increase their social distance from less powerful individuals . . . (7) decrease their social distance from less powerful people.

3-33. When in disagreement with adults, young people should defer to elders.

(1) Strongly agree . . . (7) Strongly disagree

3-35. I believe that power should be:

(1) concentrated at the top . . . (7) shared throughout the organization

Institutional Collectivism

3-7. I believe that in general, leaders should encourage group loyalty even if individual goals suffer.

(1) Strongly agree . . . (7) Strongly disagree.

3-12. I believe that the economic system in this society should be designed to maximize:

(1) individual interests . . . (7) collective interests.

3-36. In this society, most people prefer to play:

(1) individual sports . . . (7) team sports.

3-37. I believe that:

(1) group cohesion is better than individualism . . . (7) individualism is better than group cohesion

Gender Egalitarianism

3-17. I believe that boys should be encouraged to attain a higher education more than girls.

(1) Strongly agree . . . (7) Strongly disagree.

3-22. I believe that there should be more emphasis on athletic programs for:

(1) boys . . . (7) girls.

3-26. I believe that this society would be more effectively managed if there were:

(1) many more women in positions of authority than there are now . . . (7) many less women in positions of authority than there are now.

3-38. I believe that it should be worse for a boy to fail in school than for a girl to fail in school.

(1) Strongly agree . . . (7) Strongly disagree.

3-39. I believe that opportunities for leadership positions should be:

(1) more available for men than for women . . . (7) more available for women than for men

Uncertainty Avoidance

3-1. I believe that orderliness and consistency should be stressed, even at the expense of experimentation and innovation.

(1) Strongly agree . . . (7) Strongly disagree.

3-16. I believe that a person who leads a structured life that has few unexpected events:

(1) has a lot to be thankful for . . . (7) is missing a lot of excitement.

3-19. I believe that societal requirements and instructions should be spelled out in detail so citizens know what they are expected to do.

(1) Strongly agree . . . (7) Strongly disagree.

3-24. I believe that society should have rules or laws to cover:

(1) almost all situations . . . (7) very few situations.

3-25. I believe that leaders in this society should:

(1) provide detailed plans concerning how to achieve goals . . . (7) allow the people freedom in determining how best to achieve goals.

(Appendices continue)

Appendix B
Ancillary Tables and Results

Table B1
Correlations of Key Variables

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
2	0.20																
3	0.20	0.06															
4	-0.33	0.11	-0.31														
5	-0.19	0.06	-0.59	0.51													
6	-0.21	-0.04	-0.49	0.36	0.43												
7	-0.30	0.08	-0.43	0.62	0.54	0.69											
8	0.21	0.04	0.49	-0.36	-0.43	-1.00	-0.69										
9	0.18	0.14	0.55	-0.10	-0.14	-0.26	-0.17	0.26									
10	0.36	0.09	0.22	-0.26	-0.20	-0.18	-0.26	0.18	0.10								
11	0.45	0.13	0.39	-0.39	-0.30	-0.32	-0.38	0.32	0.27	0.52							
12	0.42	0.08	0.36	-0.42	-0.32	-0.30	-0.42	0.30	0.16	0.62	0.57						
13	0.49	0.10	0.55	-0.55	-0.41	-0.45	-0.54	0.45	0.38	0.49	0.71	0.80					
14	0.25	0.08	0.00	-0.25	-0.06	-0.16	-0.21	0.16	0.02	0.29	0.33	0.47	0.47				
15	0.16	0.48	0.13	0.23	0.14	-0.08	0.17	0.08	0.28	0.10	0.15	0.16	0.20	0.16			
16	0.26	0.03	0.31	-0.34	-0.22	-0.32	-0.38	0.32	0.29	0.20	0.31	0.32	0.44	0.14	0.06		
17	0.19	0.12	-0.07	0.03	0.08	0.06	0.09	-0.06	-0.04	0.05	0.09	0.02	0.00	0.06	0.07	-0.02	
18	0.29	0.15	0.08	-0.10	-0.07	-0.08	-0.05	0.08	0.01	0.15	0.19	0.12	0.13	0.10	0.05	0.03	0.25
19	0.09	0.09	0.00	0.01	0.00	0.01	0.01	-0.01	0.00	-0.01	0.02	0.00	0.00	0.01	0.04	-0.02	0.20
																	0.07

Note. Variables: (1) Reading achievement, (2) Parent rating of past literacy skills, (3) Log gross domestic product (GDP) per capita, (4) Gini index, (5) Clustering students across schools by past achievement, (6) Power distance, (7) In-group collectivism, (8) Gender egalitarianism, (9) Uncertainty avoidance, (10) Socioeconomic status (SES), (11) Home education resources, (12) Class mean SES, (13) Class mean home education resources, (14) Class mean attitude toward reading, (15) Class mean past achievement, (16) Availability of school resources, (17) Students' attitude towards reading, (18) Students' reading self-concept, (19) Girl.

(Appendices continue)

Table B2
Number of Classes, Mean Number of Students Per Class, and Proportion of Missing Data in Each Country

Country	Number of classes	Mean students per class	% Missing
Argentina	138	23.9	12
Belize	139	20.9	10
Bulgaria	170	20.4	5
Canada	409	20.2	10
Colombia	196	26.2	9
Cyprus	150	20.0	8
Czech Republic	141	21.4	7
France	221	16.6	10
Germany	393	19.4	10
Greece	145	17.2	8
Hong Kong	147	34.4	12
Hungary	216	22.0	7
Iceland	242	15.2	11
Iran	282	26.3	6
Israel	147	27.0	6
Italy	184	19.0	6
Kuwait	265	26.9	13
Latvia	141	21.4	7
Lithuania	146	17.6	8
Moldova	150	23.6	4
Netherlands	195	21.1	7
New Zealand	173	14.5	9
Norway	199	17.4	5
Romania	167	21.7	5
Russian Federation	206	19.9	6
Singapore	196	35.7	4
Slovak Republic	176	21.6	3
Slovenia	155	19.0	7
Sweden	344	20.9	6
Turkey	154	33.3	8
Macedonia	159	23.6	13
England	132	23.9	14
Scotland	136	20.0	8
Overall	6414	22.0	8

(Appendices continue)

Table B3

Country Means and Standard Deviations of Key Variables

Country name	Reading achievement		Class mean SES		Class mean home education resources		Class mean parent attitude toward reading		Class mean attitude towards reading		Class mean parent rating of past literacy skills	
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
Argentina	417.85	96.02	-0.45	0.43	-0.69	0.44	-0.20	0.20	-0.23	0.24	-0.03	0.31
Belize	326.26	106.42	-0.59	0.60	-0.75	0.46	-0.11	0.31	-0.25	0.33	-0.13	0.41
Bulgaria	550.37	82.92	-0.16	0.73	-0.25	0.79	0.10	0.67	0.21	0.37	0.08	0.62
Canada	544.31	71.21	0.54	0.44	0.58	0.38	0.11	0.32	0.00	0.34	0.27	0.33
Colombia	422.23	80.74	-0.67	0.69	-1.13	0.46	-0.02	0.34	0.16	0.29	-0.01	0.47
Cyprus	493.56	82.15	0.03	0.41	-0.03	0.24	0.07	0.22	0.02	0.29	-0.07	0.22
Czech Republic	536.81	63.96	0.08	0.31	0.44	0.33	0.07	0.26	-0.28	0.34	-0.61	0.25
France	525.40	70.48	-0.01	0.46	0.30	0.41	-0.14	0.29	0.04	0.28	0.16	0.25
Germany	538.89	66.84	0.38	0.29	0.39	0.32	-0.15	0.35	-0.07	0.31	-0.37	0.26
Greece	524.45	72.68	-0.06	0.65	-0.06	0.42	0.25	0.33	0.18	0.39	0.34	0.28
Hong Kong	527.75	63.38	-0.56	0.50	-0.43	0.41	-0.31	0.21	-0.22	0.28	0.45	0.16
Hungary	543.54	65.27	0.16	0.44	0.40	0.50	0.40	0.37	-0.17	0.39	-0.69	0.29
Iceland	511.72	75.43	0.28	0.48	0.78	0.26	0.00	0.29	0.18	0.34	-0.18	0.27
Iran	414.73	92.27	-1.11	0.58	-1.16	0.64	-0.01	0.42	0.06	0.43	-0.14	0.72
Israel	507.67	94.07	0.11	0.36	0.11	0.33	-0.06	0.22	-0.08	0.31	0.42	0.30
Italy	541.00	71.01	-0.25	0.38	-0.12	0.37	-0.05	0.32	-0.17	0.35	-0.15	0.29
Kuwait	395.90	89.48	0.01	0.13	-0.34	0.35	-0.26	0.17	-0.04	0.30	-0.33	0.26
Latvia	544.74	60.72	0.21	0.36	0.27	0.41	-0.08	0.24	-0.22	0.30	0.21	0.33
Lithuania	543.53	63.99	0.28	0.39	-0.16	0.41	-0.08	0.26	-0.04	0.33	0.09	0.33
Moldova	491.40	74.74	-0.19	0.48	-0.85	0.54	-0.24	0.36	0.17	0.36	-0.34	0.50
Netherlands	554.38	57.53	-0.03	0.31	0.35	0.30	-0.09	0.23	-0.24	0.35	-0.24	0.23
New Zealand	528.86	93.46	0.40	0.51	0.58	0.41	0.20	0.36	-0.01	0.39	0.11	0.27
Norway	499.44	82.11	0.58	0.48	0.80	0.27	0.32	0.34	-0.08	0.32	-0.02	0.25
Romania	511.80	89.01	-0.38	0.52	-0.68	0.63	-0.28	0.56	0.29	0.32	-0.23	0.45
Russian Federation	528.58	67.83	0.35	0.43	-0.07	0.54	-0.03	0.38	0.16	0.35	-0.26	0.59
Singapore	528.29	91.02	-0.04	0.59	0.42	0.48	-0.20	0.21	0.12	0.32	0.79	0.41
Slovak Republic	518.86	70.02	0.12	0.41	0.07	0.47	0.28	0.32	-0.20	0.34	-0.65	0.27
Slovenia	501.20	71.91	0.04	0.36	0.06	0.34	0.14	0.26	0.05	0.35	0.13	0.26
Sweden	561.17	65.49	0.54	0.47	0.88	0.37	0.32	0.34	0.01	0.32	0.18	0.27
Turkey	449.01	86.64	-0.51	0.50	-1.03	0.59	-0.28	0.45	0.30	0.28	-0.16	0.47
Macedonia	441.43	102.24	-0.35	0.47	-0.37	0.39	0.12	0.35	0.36	0.28	0.34	0.32
England	553.42	85.27	0.10	0.33	0.55	0.35	0.03	0.26	-0.27	0.36	0.18	0.18
Scotland	528.54	84.57	0.14	0.35	0.36	0.38	0.07	0.27	-0.14	0.42	-0.08	0.20

Note. SES = socioeconomic status.

Received March 29, 2013

Revision received March 25, 2014

Accepted March 26, 2014 ■

To What Extent Do Teacher–Student Interaction Quality and Student Gender Contribute to Fifth Graders’ Engagement in Mathematics Learning?

Sara E. Rimm-Kaufman
University of Virginia

Alison E. Baroody
San Francisco State University

Ross A. A. Larsen
Virginia Commonwealth University

Timothy W. Curby
George Mason University

Tashia Abry
Arizona State University

This study examines concurrent teacher–student interaction quality and 5th graders’ ($n = 387$) engagement in mathematics classrooms ($n = 63$) and considers how teacher–student interaction quality relates to engagement differently for boys and girls. Three approaches were used to measure student engagement in mathematics: Research assistants observed engaged behavior, teachers reported on students’ engagement, and students completed questionnaires. Engagement data were conducted 3 times per year concurrent with measures of teacher–student interaction quality. Results showed small but statistically significant associations among the 3 methods. Results of multilevel models showed only 1 significant finding linking quality of teacher–student interactions to observed or teacher-reported behavioral engagement; higher classroom organization related to higher levels of observed behavioral engagement. However, the multilevel models produced a rich set of findings for student-reported engagement. Students in classrooms with higher emotional support reported higher cognitive, emotional, and social engagement. Students in classrooms higher in classroom organization reported more cognitive, emotional, and social engagement. Interaction effects (Gender \times Teacher–student interaction quality) were present for student-reported engagement outcomes but not in observed or teacher-reported engagement. Boys (but not girls) in classrooms with higher observed classroom organization reported more cognitive and emotional engagement. In classrooms with higher instructional support, boys reported higher but girls reported lower social engagement. The discussion explores implications of varied approaches to measuring engagement, interprets teacher–student interaction quality and gender findings, and considers the usefulness of student report in understanding students’ math experiences.

Keywords: engagement, teacher–student interactions, mathematics, classrooms, fifth grade

There have been remarkable shifts in mathematics education in the past 2 decades. The stance advanced by the National Council for Teachers of Mathematics (NCTM, 2000) and codified in the U.S. Common Core State Standards Initiative (CCSSI, 2014) describes learning mathematics as a dynamic, exploratory process focused on creating opportunities for students to develop a con-

ceptual understanding of mathematics. Students are expected to identify and describe patterns in mathematics, participate in conversations about mathematical problem solving, use mathematics to think and reason, and justify their mathematical thinking. The new emphasis contrasts with a traditional view describing mathematics education as a static set of facts, concepts and procedures to

This article was published Online First July 28, 2014.

Sara E. Rimm-Kaufman, Curry School of Education & Center for Advanced Study of Teaching and Learning, University of Virginia; Alison E. Baroody, Department of Child and Adolescent Development, San Francisco State University; Ross A. A. Larsen, Department of Foundations of Education, Virginia Commonwealth University; Timothy W. Curby, Department of Applied Developmental Psychology, George Mason University; Tashia Abry, T. Denny Sanford School of Social and Family Dynamics, Arizona State University.

The research reported here is based upon work supported by the National Science Foundation under Grant DRL-0814872. The work was also supported by training grants from the Institute of Education Sciences, U.S. Department of Education, through Grant R305B040049 and Grant

R305B060009 to the University of Virginia. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation or U.S. Department of Education. We gratefully acknowledge the contributions of Sandra Christenson, Julia Thomas, Michelle Ko, Claire Cameron, Eileen Merritt, Abigail Moncrief, Jennifer Williams, and the administrators, teachers, families and students in our collaborating school district.

Correspondence concerning this article should be addressed to Sara E. Rimm-Kaufman, Curry School of Education and Center for Advanced Study of Teaching and Learning, University of Virginia, Ruffner Hall, Emmet Street South, Charlottesville, VA 22904. E-mail: serk@virginia.edu

be learned and memorized (Henningsen & Stein, 1997; Hiebert & Grouws, 2007; Schoenfeld, 1992).

Standards-based mathematics education heightens the demand for students to be actively engaged in learning. Imagine the challenge from the perspective of a fifth grade teacher striving to produce high levels of engagement in her classroom. The teacher may have been handed clear guidelines on what to teach; for instance, CCSSI guidelines emphasize multiplication and division of fractions and fluency in operations with multidigit whole numbers and decimals to the hundredths (CCSSI, 2014). However, how to teach in a way that fully engages students is much less clear. Fifth graders are experienced students and often know how to comply and appear engaged in learning. However, learning mathematics requires much more than simply the appearance of engagement. Students need to feel engaged in math learning for the instruction to take hold. Students need to pay attention, become interested in the mathematical ideas, and even work together with others on math problems. The heightened demands establish the need for research that identifies conditions that foster student engagement.

Despite the large body of research on engagement, few studies focus exclusively on math, little work measures teacher behaviors and student engagement concurrently, and work using student-report data is scarce (Christenson, Reschly, & Wylie, 2012; Fredricks, Blumenfeld, & Paris, 2004). Finn and Zimmer (2012) have stated a need for research on classroom contexts that support and threaten student engagement: “A package of assessments for this purpose would involve observations of students in the school setting, observations of teacher-student interactions (with specific foci), and reactions from students themselves” (p. 125).

The present study addresses this stated need. Five unique contributions stand out. First, we gather student engagement data based on observational, teacher-report, and student-report measures. Second, we measure teacher-student interaction quality and engagement concurrently to understand the temporal coupling between teachers’ behaviors and students’ experience. Third, we gather data in mathematics classrooms only, whereas most research on teacher-student interaction quality is not content specific. Fourth, we consider various facets of teacher-student quality. We examine teacher sensitivity and supportiveness, the approach to behavior management, and opportunities for higher order thinking and back-and-forth conversation between teachers and students. Fifth, we examine the extent to which classroom conditions are equivalently important for girls and boys. Thus, our goal is to identify immediate classroom conditions that enhance and diminish engagement for fifth grade boys and girls. The work was designed to provide basic research insights for mathematics educators concerned with leveraging teacher-student interaction quality to improve engagement in math learning.

Theoretical Perspective

The work was guided by an integrative framework of motivation (Skinner, Kindermann, Connell, & Wellborn, 2009) that explains how characteristics of children’s contexts contribute to self-systems and self-perceptions, which lead to action (engagement or disaffection) and ultimately to social, emotional, and academic outcomes (Skinner et al., 2009). Skinner et al. (2009) defined children’s contexts as settings composed of peers, teachers, family

members and others with whom children engage in social interactions and activities. Self-systems and self-perceptions refer to children’s beliefs, cognitive appraisals, and perceptions of themselves that develop in children as a result of their past experiences, mold children’s interpretation of their experiences, and play an important role in motivating children’s behavior. Action refers to engagement versus disaffection; each reflects an outward signal of motivational state and describes the quality of children’s interactions with their physical and social world. The outcomes include social, cognitive and personality development.

The integrative framework of motivation describes motivation as a dynamic, developing characteristic that is sensitive to contexts external to the child (e.g., interactions with teachers). Further, the framework introduces the utility of measuring a child’s engagement at a particular point in time as one way to “capture the target definitional manifestations of motivation—namely, energized, directed, and sustained action” (Skinner et al., 2009, p. 225). This view applies to students in elementary math classrooms. Most fifth grade students do not come to math class as “engaged” or “disengaged.” Students’ engagement in math class varies depending on their interactions with teachers, peers, and materials (Connell & Wellborn, 1991; Skinner & Belmont, 1993). Further, students’ engagement varies across days, weeks, and months—a student who appears engaged in mathematics instruction one day may be less engaged in math class on a day 1 full month later.

We apply the integrative framework of motivation to understand day-to-day interactions between teachers and students. Each year of math instruction is composed of daily experiences that vary in quality and accrue to create a cumulative experience for students. We disaggregate the year of math instruction by sampling specific days and assessing the immediate correspondence between teachers’ interactions with students and students’ engagement. The work is situated at an important point developmentally; fifth graders are capable of reflecting upon and reporting their engagement in learning, and fifth grade marks a turning point when boys begin to outperform girls in mathematics (Robinson & Lubinski, 2011).

Engagement in Learning

Engagement has been described as “the glue, or mediator, that links important contexts—home, school, peers, and community—to students and, in turn, to outcomes of interest” (Reschly & Christenson, 2012, p. 3). Existing research establishes that engagement is critical for learning and that engagement forecasts school success. Students who stay on task, attend to learning goals, and participate actively in the learning experience show better academic achievement in elementary school (Fredricks et al., 2004; Greenwood, Horton, & Utley, 2002; Hughes & Kwok, 2007; Ladd, Birch, & Buhs, 1999; Ponitz, Rimm-Kaufman, Grimm, & Curby, 2009; Reyes, Brackett, Rivers, White, & Salovey, 2012; Tucker et al., 2002).

Definitions of engagement vary considerably; a three-part definition of engagement that includes behavioral, cognitive, and emotional engagement is most prevalent (Reschly & Christenson, 2012; Fredricks et al., 2004). Behavioral engagement refers to paying attention, completing assigned work, participating in teacher-sanctioned learning opportunities, and showing an absence of disruptive behaviors. Cognitive engagement refers to a willingness to exert effort to understand content, work through difficult

problems, and manage and direct their attention toward the task at hand. Emotional engagement refers to feelings of connection to content, interest in learning, and enjoyment of solving problems and thinking about content (Fredricks et al., 2004). A fourth construct, social engagement, is also fundamental. Social engagement (termed "task-related interaction" by Patrick, Ryan, & Kaplan, 2007) refers to students' day-to-day social exchanges with peers that are tethered to the instructional content. Standards-based math instruction emphasizes activities involving small groups of students and mathematical discourse among students (Fuson, Kallchman, & Bransford, 2005; NCTM, 2000).

Although conceptualizations of engagement vary, there are four common themes: (a) Engagement is a critical mediator for learning; (b) engagement is multifaceted with behavioral, cognitive, emotional, and social elements; (c) different sources of data are necessary depending on the type of engagement measured; and (d) students show a decrease in engagement in learning as they progress from elementary school into the middle school years (Furrer & Skinner, 2003; Marks, 2000; Reschly & Christenson, 2012; Reyes et al., 2012). Engagement theories describe dynamics within the engagement system (e.g., emotional engagement stimulates behavioral engagement) and outside of the engagement system (e.g., social context contributes to behavioral engagement; Reschly & Christenson, 2012; Skinner et al., 2009). We focus on a factor outside of the engagement system, teacher-student interaction quality, and consider the extent to which it contributes to behavioral, cognitive, emotional, and social engagement.

Teacher-Student Interactions

Teachers' interactions with students vary in quality and have appreciable effects on math achievement outcomes (Martin, Anderson, Bobis, Way, & Vellar, 2012; Reyes et al., 2012). Teacher-student interactions are malleable features of classroom environments and have been the focus of national efforts to raise mathematics achievement (Pianta & Hamre, 2009; Rimm-Kaufman & Hamre, 2010).

Quality of Teacher-Student Interactions

Teacher-student interaction quality can be described in relation to three domains: emotional, organizational, and instructional support (Pianta & Hamre, 2009). Emotional support refers to the teachers' connection to and responsiveness toward students, awareness of students' individual differences and needs, and willingness to incorporate students' point of view into learning activities. Classroom organization refers to the teachers' tendency to use proactive rather than reactive supports to foster classroom routines and guide classroom behavior, use instructional approaches that make learning objectives clear, and use a variety of modalities to engage students in learning. Instructional support refers to the presence of feedback loops in teacher-student communication and provision of opportunities to engage in higher order thinking and learn new language and vocabulary (Pianta, La Paro, & Hamre, 2008).

Research links teachers' emotional support (i.e., positive classroom social climate, teacher sensitivity toward students) to enhanced engagement in kindergarten (Rimm-Kaufman et al., 2002) and third grade classrooms (NICHD Early Child Care Research

Network, 2005). Teacher efforts to reinforce prosocial behavior in sixth and seventh grade contribute to enhanced behavioral and social engagement (Matsumura, Slater, & Crosson, 2008). Meta-analytic work has demonstrated associations between positive teacher affect and engagement and between negative teacher affect and disengaged behavior (Roorda, Koomen, Spilt, & Oort, 2011). Engagement plays a mediational role linking emotional support to achievement in both upper elementary (Reyes et al., 2012) and middle school grades (Voelkl, 1995).

High quality classroom organization has been linked to engagement in kindergarteners, first graders (Ponitz, Rimm-Kaufman, Brock, & Nathanson, 2009; Rimm-Kaufman, Curby, Grimm, Nathanson, & Brock, 2009), and third graders (NICHD Early Child Care Research Network, 2005). Teachers who establish clear routines in the fall appear to increase the self-regulated behavior of their students throughout the school year (Bohn, Roehrig, & Pressley, 2004; Cameron, Connor, & Morrison, 2005). Third graders in classrooms with higher levels of productivity and more opportunities to engage in academic instruction spend more time behaviorally engaged (NICHD Early Child Care Research Network, 2005).

Instructionally rich learning environments are also likely to support engagement. The presence of authentic learning experiences (i.e., provision of interesting questions and opportunities for in-depth learning) relates to increased student engagement in math learning in elementary and middle school years (Marks, 2000). In sixth and seventh grade classrooms, teachers who asked students challenging questions and encouraged students to explain the evidence behind their statements in classroom discussions enhanced the quality of the discourse (Matsumura et al., 2008), and thus, participatory engagement. Middle school teachers who were observed showing high expectations for student work, monitoring student progress and providing scaffolding, and challenging student thinking produced higher levels of student engagement (Raphael, Pressley, & Mohan, 2008; Woodward et al., 2012).

Gender

Student gender has been linked to engagement; boys show lower levels of behavioral and emotional engagement than girls in elementary and middle school (Kindermann, 2007; Marks, 2000). Most work demonstrating gender differences in engagement generalizes across content areas and is not specific to math. The fifth grade year appears to be an important time to compare the engagement of boys and girls in math because it marks an inflection point in achievement. From kindergarten to fifth grade, math achievement disparities between boys and girls increase, with boys showing more achievement growth than girls. In middle school, the reverse is true, and girls show larger achievement increases than boys (Robinson & Lubienski, 2011). Gender disparities in engagement and achievement warrant further investigation in math classrooms.

Simply comparing engagement between boys and girls does not fully recognize the role of teacher-student interactions in the facilitation of engagement. In a study of elementary and middle school students, the higher level of engagement in girls than boys was attenuated in the presence of social support (i.e., student-reported teacher respect, feelings of safety at school, the presence of high expectations from their teachers, and opportunities to

discuss academic issues with their families; Marks, 2000). Other work has described boys as having more frequent and academically challenging interactions with their teachers than girls (National Research Council, 2001). Too little is known about how teachers' interactions with students are differentially important to boys' versus girls' concurrent engagement in fifth grade math.

Informants of Engagement

We used three categories of informants to assess engagement. Observers measured behavioral engagement; teachers reported on behavioral engagement; and students reported on their cognitive, emotional, and social engagement. Each information source provides a unique perspective. Classroom observational methods measure observable indicators of engagement, such as behavioral engagement (Brophy & Good, 1986; NICHD Early Child Care Research Network, 2005) but do not capture students' internal psychological experience. Teachers' ratings provide global indicators of students' engagement and provide cumulative reports of engagement over a year but also tap teachers' subjective beliefs about students (Gest, Domitrovich, & Welsh, 2005; Mashburn, Hamre, Downer, & Pianta, 2006). Student-report methods provide students' own perspective of their psychological experience and may be better for measuring intrapsychic experiences; however, student-report methods may be sensitive to social desirability bias. We included three informants because each reporter provides a unique perspective on students' engagement.

Other Factors

Several student attributes with theoretical or empirical links to mathematics engagement were included as covariates. Age was included because of its association with self-regulatory abilities in school settings (Bronson, 2001). Eligibility for free or reduced priced lunch (FRPL) was used as an indicator of low income; elementary school-aged children living in impoverished environments experience confrontation with chronic stressors linked to lower self-regulatory abilities and engagement (Evans & English, 2002; Evans & Rosenbaum, 2008). Initial achievement was included because of associations between higher math achievement and emotional engagement, and because rate of growth in math learning differs for students with preexisting academic difficulty compared to typical students (Bodovski & Farkas, 2007; Crosnoe et al., 2010; Dotterer & Lowe, 2011). Self-efficacy in math, defined as students' perception of their capacity to learn or perform in math, was included because of established links to engagement (Linnenbrink & Pintrich, 2003; Schunk & Pajares, 2005). Time of year was included because of changes in student experience over the year (Curby, Rimm-Kaufman, & Abry, 2013).

The Present Study

We address three questions. First, to what extent do observationally based, teacher-reported, and student-reported measures of engagement show concordance and discordance? We hypothesized stronger associations within informants than among informants. We expected that varied approaches to measurement would provide different lenses on engagement. Second, to what extent do the quality of teacher-student interactions and student gender contrib-

ute to engagement? We expected higher engagement among girls than boys and expected higher quality teacher-student interactions to relate to higher engagement. Third, does the quality of teacher-student interactions predict student engagement differentially for boys and girls? We expected higher quality teacher-student interactions would be more important for engaging boys than girls.

Method

Participants

All schools were located in a single suburban district in a Mid-Atlantic state. Schools and fifth grade teachers were recruited by the research team through in-person meetings with principals and teachers. Response rates were 83% and 79% for schools and teachers, respectively. The selected schools ($N = 20$) were socio-economically and linguistically diverse; 33% of students qualified for FRPL, and 31% were English language learners (ELL). Sixty-three fifth grade mathematics teachers participated. Teachers had, on average, 12.49 years of experience (range = 1–38). Most teachers were Caucasian ($n = 48$); five were Hispanic, one was African American, one was Native American, and two were multiracial. Six teachers did not report their race/ethnicity. All teachers held bachelor's degrees; 38 had master's degrees. All teachers reported having a full state certification. Teachers received financial remuneration for participating.

Fifth grade students were recruited via mailings sent home to all parents by participating teachers in the fall of students' fifth grade year. Family recruitment practices followed customary district procedures for family communication, involving translation into seven commonly spoken languages. Parents of 479 students signed consent forms and received gift certificates for participating.

Approximately five students per classroom (mode = 5) were selected from the 479 consented students, resulting in the sample of 387. Selection was conducted randomly for each classroom bounded by two constraints: (a) maintenance of equal number of girl and boy participants, and (b) demographic match to the whole school (based on ethnicity, FRPL, and ELL percentages). The final sample of student participants ($n = 387$; 203 girls) were 10.47 years old ($SD = 0.37$) in September 2010. School records showed that 21% of students qualified for FRPL (income of \$40,793 for a family of four, roughly below 180% of the federal poverty guideline). Parent-report questionnaires (described below) showed that 55% of students spoke primarily English at home, 28% spoke a non-English language (22 different languages reported), and 17% had missing data. Of the 321 parents reporting mothers' education, 7.2% did not have a high school diploma, 21.5% had a high school diploma, and 71.3% had an associate's degree or above.

Procedures

The research team began by conducting extensive, iterative pilot work in 2009–2010 with 33 fifth grade students and six fifth grade teachers. Existing, well-validated measures were used to measure engagement, when available. The research team reviewed existing measures and found that typical engagement measures were not necessarily well suited for fifth graders, math, or reflections on one specific day of class. Pilot work was conducted to adapt existing

observational and student-report engagement measures to meet high levels of rigor.

Data were gathered from five sources: (a) district data, (b) parent-report questionnaire, (c) classroom observations, (d) student-report questionnaires, and (e) teacher-report questionnaires. All data were gathered while students were enrolled in fifth grade except for initial student achievement data, which was gathered by the district as part of the end-of-the-year fourth grade standardized testing. Data were gathered between May 2010 and May 2011, as shown in Table 1.

Data gathered by the district were used to measure student FRPL and initial achievement. Parents completed a demographic questionnaire at fall recruitment to describe student and sample characteristics (e.g., gender, age, mother's education). Pairs of research assistants conducted classroom observations in math classes at three times during the school year, corresponding to three windows (Window 1: late September to late November; Window 2: late November to mid-February; and Window 3: late February to late April). At each observation, one research assistant videotaped the classroom to gather teacher-student interaction quality data and a second research assistant live-coded student engagement (as described below). The research assistants administered engagement questionnaires to student participants immediately after each observation. After the fall observation only, students completed a questionnaire about their math self-efficacy. In spring, teachers completed a teacher demographic questionnaire and questionnaires about each participant to measure student engagement in mathematics learning.

All classroom visits were scheduled and conducted following a specific protocol. Research assistants scheduled classroom observations of teachers and students on days that teachers deemed "typical days" of math instruction. Observations were scheduled for 3 different days during the school year to sample typical practices from over 3 hr of observation. Classroom observations were conducted for the full length of the math lesson ($M = 63$ min, range 15 to 135). One research assistant began videotaping the classroom prior to the transition to math instruction and ended at the end of the math lesson. The second research assistant (child

observer) conducted two 4-min observations of engagement for each child participant during the same time in which the teacher was observed. Child observers followed a protocol in which they would watch one student for 4 min, complete ratings, watch a second student for 4 min, complete ratings, and so on, until all student participants had been observed and rated once. Then, the child observer would cycle through the student participants a second time, observing each student for 4 min and completing ratings again. Child observations resulted in 24 min of observed engagement per child. Child observers made efforts so that students were unaware that they were being observed. When the math lesson observation was complete, the research assistants distributed student-report engagement questionnaires to student participants to measure their engagement in mathematics on that specific day. All classroom videotapes were sent to the laboratory for subsequent coding.

Measures

District data.

Student demographic data. District records were used to determine eligibility for FRPL.

Initial achievement. The paper version of the state standardized test, the Standard of Learning (SOL), was administered by the district to assess fourth grade mathematics achievement (Virginia Department of Education [VDOE], 2008). The test was composed of 50 multiple choice items tapping students' procedural knowledge and conceptual understanding of four skill categories: (a) number and number sense, (b) computation and estimation, (c) measurement and geometry, and (d) probability, statistics, patterns, functions and algebra (VDOE, 2010). The state computed the total number of items correct and converted the number to a scaled score ranging from 0 to 600. A scaled score of 400 indicates pass/proficient, and 500 indicates pass/advanced. The *Virginia Standards of Learning Technical Report* (VDOE, 2008) describes test development, calibration, and validity. Test items were developed through a collaborative process among Virginia educators,

Table 1
Timeline for Data Collection From Five Data Sources

Data source and type	May 2010	Sept.-Nov. 2010	Nov. 2010–Feb. 2011	Feb.–April 2011	April–May 2011
District data					
Student demographic information	x				
Initial achievement test (4th grade)	x				
Parent-report questionnaire			x		
Student demographic questionnaire			x		
Classroom observations					
Teacher-student interaction quality	x		x		x
Observed engagement	x		x		x
Student-report questionnaires					
Math self-efficacy	x				
Engagement in mathematics	x		x		x
Teacher-report questionnaires					x
Engagement questionnaire	x				x
Teacher demographic questionnaire					x

VDOE, Educational Testing Service, Pearson, and content experts based upon test blueprints. Calibration was established using Rasch modeling and the partial credit model. Test validity was established by gathering empirical evidence supporting the face validity, intrinsic rational validity, content validity, and construct validity (VDOE, 2008). Students deemed not proficient in English were administered the Plain English Math version that equates to the standard math assessment.

Parent-report questionnaires.

Student demographic information. Parents completed customized questionnaires to describe sociodemographic characteristics. Gender was coded as 1 for female. Child age in September 2010 was computed based on birthdate. Parents reported primary language spoken at home and level of maternal education.

Classroom observations.

Teacher-student interaction quality. Quality of teacher-student interactions was assessed using the Classroom Assessment Scoring System (CLASS; Pianta et al., 2008). There are 10 measured dimensions that correspond to three domains (emotional support, classroom organization, and instructional support). Emotional support was measured using dimensions of positive climate, negative climate, teacher sensitivity, and regard for student perspectives ($\alpha = .83$). Positive climate referred to a positive emotional tone among teachers and students and referred to respect, enthusiasm, and evidence of enjoyment. Negative climate (reversed for analysis) tapped teachers' evidence of sarcasm, anger, aggression and/or harshness. Teacher sensitivity measured evidence of the teacher providing comfort, reassurance and encouragement in relation to students' academic and social needs. Regard for students' perspectives referred to situations in which teachers' choice of classroom activities demonstrated emphasis on students' motivation, interests, and point of view.

Classroom organization was assessed using scales of behavior management, productivity, and instructional learning formats ($\alpha = .82$). Behavior management measured teachers' use of effective methods to prevent students' behavior problems and redirect students, as needed. Productivity referred to the teachers' use of instructional time and routines enabling appropriate learning opportunities for students. Instructional learning formats referred to teachers' use of materials and activities to facilitate learning opportunities.

Raters assessed three dimensions in relation to instructional support for learning ($\alpha = .72$). Concept development measured the teachers' use of strategies to promote students' higher order thinking. Quality of feedback assessed specificity of teachers' verbal interaction pertaining to student work, ideas and comments (e.g., did teacher comments create communication loops between the teacher and students). Language modeling measured the extent to which teachers facilitated, encouraged and modeled students' use of advanced language. Each dimension was rated on a 7-point Likert scale. For analyses predicting observed and student-reported engagement, CLASS domains collected concurrently with the engagement outcome were used in models. For analyses predicting teacher-reported student engagement, mean levels of domains were calculated for each teacher across the three observation windows.

Prior to CLASS training, coders read manuals and additional readings and conducted practice observations. Training involved the time equivalent of a 2-day small group interactive training

followed by paired observations with an expert. Reliability tests involved rating ten 15-min segments for CLASS. Ratings were compared to a gold standard, prepared by the instruments' authors. To be considered reliable, each coder's responses had to be within 1 scale point of the gold standard on 80% of the responses. Reliability exceeded these levels prior to data collection. Calibration involved independent coding (once or twice per month) in a small group session followed by reliability checks and discussion of coding rationales plus double coding of more than 10% of tapes selected randomly. In addition, master coders conducted audits by coding one tape coded by each coder every 12 weeks. Measures of teacher-student interaction quality were based on two segments (minutes 0–15 and 30–45).

Observed behavioral engagement. Research assistants assessed student engagement using time-sampling and global rating systems adapted from the NICHD Early Childcare Research Network (2005) Classroom Observation Scale (COS). The COS had been used in fifth grade classrooms (Pianta et al., 2008) but required honing, revised documentation, and testing, all of which were conducted during the pilot year. The time sampling measure was a low-inference measure and required an observer to note the presence or absence of disengagement in 1-min intervals. Disengagement included wandering, looking away from instructional opportunities, behaviors disruptive to learning, and similar behaviors listed in a manual. Students were observed and coded for four consecutive 1-min intervals, twice during each math lesson. Observed on-task behavior was calculated as minutes observed minus minutes of disengagement.

The global rating was a high-inference measure composed of three rating scales: (a) participation in learning opportunities (e.g., duration and interest of involvement), (b) disruptive behavior (reversed; e.g., excessive out-of-turn talking, sustained noise), and (c) self-reliance (e.g., self-management of materials and responsibilities). Research assistants took notes related to global codes during the 4-min time-sampling period. Then, the research assistant used the notes and a scoring rubric to rate behavior from 1 (low) to 7 (high) on each scale.

Reliability training involved following a protocol with a four-phase process (preparation, training, reliability, and ongoing calibration) to attain and maintain reliability. The process was comparable to that described for the CLASS. Reliability values prior to data collection (based on eight segments) and during monthly calibration (based on eight segments), respectively, showed an intraclass correlation of .65 and .95 for the time sampling measure and 75% and 90% within one match for the global rating. Initial values were lower than desired, but later tests of reliability indicated substantial improvements. Research assistants conducted paired coding during initial visits until reliability improved. The time sampling score and global ratings of behavioral engagement were correlated ($r = .70$). All four scores were included in a confirmatory factor analysis to create a factor score.

Student-report questionnaires.

Students' feelings of math efficacy. The Academic Efficacy subscale of the Patterns of Adaptive Learning Scale (Midgley et al., 2000) was used to measure students' perception of their competence. The subscale was modified to apply to a math context and piloted and validated in a sample of 39 students (pilot $\alpha = .89$). Students rated items such as, "I'm certain I can master the skills taught in math this year," and "I can do almost all of the work in

math if I work at it" on a scale from 1 (*almost never*) to 4 (*all the time*). The five items were averaged to create a composite value of math efficacy where higher values indicated higher efficacy ($\alpha = .81$).

Student-reported engagement in math class. The student-reported measures of cognitive and emotional engagement were developed based upon measures created by Meece (2009); Kong, Wong, and Lam (2003); Rowley, Kurtz-Costes, Meyer, and Kizzie (2009); and Skinner and Belmont (1993) to assess students' report of cognitive and emotional engagement in relation to a specific day and the math context. Development and piloting of student-report measures of cognitive and emotional engagement involved selection of items by two mathematics education experts, adjustment of wording to measure a single day in math, and review by a research team. This process was followed by piloting the measure with students immediately after their math instruction. Researchers observed the students and then distributed measures. Students rated their engagement and identified confusing items. Research assistants engaged in conversations about cognitive and emotional engagement with selected students for content validation. This process was conducted serially following protocols.

The student-report measure of social engagement developed and used by Patrick et al. (2007) was adopted in its existing form (with the only modification involving addition of the phrase "in math class"). The measure of social engagement was composed of five items that measured the extent to which students explained academic content to one another and discussed ideas with other students in class.

Students reported on engagement in a 15-item questionnaire with a scale from 1 (*no, not at all true*) to 4 (*yes, very true*). The questionnaire was piloted in a sample of 33 fifth graders in three schools prior to use, resulting in an alpha for the complete measure of .90 and correlations exceeding .50 (p values $< .01$), with analogous measures given to teachers simultaneously. Using data from the present study ($n = 387$), a confirmatory factor analysis resulted in a three-factor solution representing subconstructs of cognitive engagement, emotional engagement, and social engagement, with alphas of .78, .91, and .74, respectively. The alpha values for cognitive and social engagement were not as high as desired; however, we chose to use these factors in subsequent analyses because of solid factor loadings and high fit indices (see Table 3). Factor scores for each subconstruct were used in analyses. Further validation stems from relatedness of the three subconstructs to other constructs. Factor scores for cognitive, emotional, and social engagement correlated .56, .67, and .49, respectively, with student-reported feelings about school.

Teacher-report questionnaires.

Teacher-reported engagement. Teachers reported on behavioral engagement in math using an eight-item version of the student engagement questionnaire used by Wu, Hughes, and Kwok (2010) and Skinner, Furrer, Marchand, and Kindermann (2008), and adapted to include the phrase, "in math class." Teachers rated each item from 1 (*strongly disagree*) to 4 (*strongly agree*). Items included "This student pays attention in math class" and "This student participates in discussion in math class." Factor analysis confirmed a one-factor solution. This is consistent with previous use of the measure as a single scale and is supported by a high internal consistency-reliability estimate ($\alpha = .92$). This factor score correlated significantly with fourth grade achievement ($r =$

.29, $p < .01$) and showed predictive validity to fifth grade achievement based on other analyses conducted as part of this study.

Teacher demographic questionnaire. Teachers reported gender, ethnicity, education, years of experience, certification, and other demographic characteristics in a questionnaire.

Analytic approach. The initial step involved the reduction of engagement data. We conducted separate confirmatory factor analyses (CFA) for each measure of engagement using Mplus 6.12 (Muthén & Muthén, 2010). The CFA utilized a priori decisions about factors drawn from theory, previous research, and item source (Kong et al., 2003; Meece, 2009; Rowley et al. 2009; Skinner & Belmont, 1993). Observed behavioral engagement and teacher-reported engagement were hypothesized to have only one factor based upon theory and study design, whereas student-reported engagement was hypothesized to have three factors. Following each CFA, a factor score was generated for use in analyses.

Observed behavioral engagement and student-reported engagement were collected three times per year, whereas the teacher-reported engagement measure was collected only once. For Question 1 only, we aggregated observed behavioral and student-reported engagement measures across the three observations, resulting in one value per student for each of the following: observed behavioral engagement, self-reported cognitive engagement, emotional engagement, and social engagement. Descriptive statistics were calculated to understand basic data patterns.

Question 1 examined concordance and discordance between observationally based, teacher-reported, and student-reported math engagement. Bivariate correlation coefficients were computed and examined. Questions 2 and 3 involved multilevel modeling to account for clustering effects (observations nested within students, students nested within teachers, and teachers nested within schools) using PROC MIXED in SAS (Version 9.2). Questions 2 and 3 used the five dependent variables (created from the abovementioned CFA): (a) observed behavioral engagement, (b) teacher-reported behavioral engagement, (c) student-reported cognitive engagement, (d) student-reported emotional engagement, and (e) student-reported social engagement. Both observed and student-reported engagement data were gathered at three time points. Instead of aggregating the data across time for these outcomes, we handled the longitudinal nature of the data via random effects in SAS PROC MIXED.

Model assumptions. Multilevel modeling assumes normality of the residuals, linear relationships between variables, no outliers, and an appropriate method for handling missing data (Kline, 2011; Little & Rubin 1987). Data were examined through residuals plots, histograms, and scatterplots. Assumptions were met for normality of the residuals and linear relationships between variables. No outliers were apparent. Roughly 5% of the data were missing for all the covariates. Analyses were conducted to determine the type of missing data via bivariate correlations and logistic regression. Considering the exhaustive nature of the covariates and the fact that missing data analyses revealed no systematic trends, the data were determined to be most likely missing at random. Subsequently, data were imputed in Mplus (Muthén & Muthén, 2010) while accounting for the clustering of the data.

Multilevel model building. Models were built incrementally in three steps. First, we created a basic four-level model that included gender and all covariates. Second, to address Question 2, we added one teacher-student interaction quality variable at a time to the basic model. Finally, to address Question 3, we generated models with one teacher-student interaction quality variable (e.g., emotional support) plus its corresponding interaction with gender (e.g., Gender \times Emotional support). The decision to include one teacher-student interaction variable at a time versus all three teacher-student interaction variables simultaneously involved additional consideration. Next, we describe the model building process and then we describe the decision about how we handled teacher-student interaction variables.

The basic model involved two observation-level variables (weeks in the year, weeks in the year squared), five student-level variables (gender, age, FRPL, initial achievement, self-efficacy), two teacher-level variables (master's degree, years of experience), and no school-level variables. To address Question 2, examining gender and teacher-student interaction quality main effects, we added one teacher-level variable (emotional support, classroom organization, or instructional support) to the basic model. To address Question 3, examining gender by teacher-student interactions, we used the basic model plus a cross-level interaction between gender and one domain of teacher-student interaction quality (Gender \times Emotional support, Gender \times Classroom organization, or Gender \times Instructional support). Each model included all the predictors that were part of the basic model. Centering was unnecessary because gender was binary at the student level. We conducted post hoc contrasts to test significance of the slopes for boys and girls separately.

This approach to model building was the same for four of five engagement outcomes. Observed behavioral engagement and the three student-reported engagement constructs were measured at three points during the year. Thus, we tested change over time as a linear term (weeks from the start of school) and curvilinear term (weeks from the start of school squared) to estimate slight curvature. Teacher-reported behavioral engagement was measured once. For that outcome, we used three-level models that excluded the observation level.

The data measured longitudinally were analyzed using a repeated measure analysis with random effects in SAS PROC MIXED. This procedure allows time to be a within-subject factor because different measurements on the same student are at different times (which we refer to as observation level). Time was tested as a main effect in the model; in other words, the model was designed to answer the question of how a student changes as time progresses (with potential linear and curvilinear effects). The student was entered as a random effect permitting inferences to be made to the entire population of students who could have been in the study. The clustering effects of classrooms and schools were handled similarly as random effects. The multilevel models all had the same general form.

Level 1: Observed Behavioral Engagement_{ijk} = π_{0ijk}
 $+ \pi_{1ijk}$ (Weeks) + π_{2ijk} (Weeks)² + e,

$$\text{Level 2: } \pi_{0ijk} = \beta_{00ij} + \beta_{10ij} (\text{Gender}) + \beta_{20ij} (\text{Age}) \\ + \beta_{30ij} (\text{FRPL}) + \beta_{40ij} (\text{Initial Achievement}) \\ + \beta_{50ij} (\text{Self - Efficacy}) + r,$$

$$\text{Level 3: } \beta_{00ij} = \gamma_{000i} + \gamma_{100i} (\text{Masters Degree})_t \\ + \gamma_{200i} (\text{Years of Experience})_t \\ + \gamma_{300i} (\text{Teacher - Student Interaction Quality Domain})_t + u,$$

$$\text{Level 4: } Y_{000i} = \eta_{0000} + \varepsilon,$$

where

$$\begin{aligned} i &= 1, \dots, 20 (\text{schools}) \\ j &= 1, \dots, 63 (\text{classrooms}) \\ k &= 1, \dots, 387 (\text{students}) \\ t &= 1, \dots, 3 (\text{time points}). \end{aligned}$$

Levels 1, 2, 3, and 4 correspond to time (observation), student, teacher, and school-levels, respectively. Weeks and Weeks² were set as fixed effects. Level 1 was omitted for the teacher-reported behavioral engagement model.

Handling of correlated CLASS domains. Correlation coefficients showed associations among the CLASS domains (emotional support, classroom organization, instructional support) with coefficients ranging from .58 to .62. Including all three domains in the same model raises multicollinearity concerns. However, analyzing each domain separately means that model results for each domain also contain information about the portion of variance shared across domains. Resolution involved a two-part approach: First, we analyzed each domain alone in separate models (keeping all covariates the same); second, we computed each model with all three domains entered simultaneously. We compared results and considered trade-offs. Results for emotional support and classroom organization were comparable regardless of analytic approach. However, in the model with all domains entered simultaneously, instructional support was negatively associated with each outcome. The negative association contradicted theory, hypotheses, and the positive relations evident in the zero order correlations. As a result, we decided to report results from models that included each CLASS domain separately, in keeping with work elsewhere (Avant, Gazelle, & Faldowski, 2011; Rudasill, Gallagher, & White, 2010).

Results

Factor Analysis

The CFA for observed behavioral engagement (based on time-sampled frequency of engagement and global engagement ratings) revealed an excellent model fit for the hypothesized one-factor solution (CFI = .99, TLI = .97, RMSEA = .09, SRMR = .02). Table 2 shows factor loadings for observer-reported engagement. We hypothesized a one-factor solution for teacher-reported behavioral engagement. The CFA was conducted and the resulting fit statistics were excellent (CFI = .95, TLI = .93, RMSEA = .04, SRMR = .04). Results are shown in Table 3. Three types of student engagement were hypothesized for the student-report measure: cognitive, emotional, and social engagement. For cognitive

Table 2
Confirmatory Factor Analysis for Observer-Reported Behavioral Engagement

Item	Standardized factor loading
Observed on-task behavior (based on frequency of engaged behavior)	0.81
Participation in learning opportunities (based on global rating)	0.92
Disruptive behavior (based on global rating)	-0.56
Self-reliance (based on global rating)	0.90

Note. CFI = .99, TLI = .97, RMSEA = .09, SRMR = .02.

engagement, two items ("I thought about other things instead of math in math class today" and "Today I only paid attention in math when it was interesting") had relatively weak loadings compared to other items but were included because all factor loadings were significant ($p < .01$) and the model fit was excellent (CFI = .96, TLI = .96, RMSEA = .03, SRMR = .04). For emotional engagement and social engagement, the factor loadings were all relatively strong, as shown in Table 4.

Descriptive Statistics and Correlations

Mean values suggest that, on average, students were highly engaged in math class regardless of informant. Students were observed as behaviorally engaged 3.28 min of the 4.0 min observed and were rated by observers as 5.66 on behavioral engagement on a 1–7 scale (based on values obtained prior to the CFA). On average, student-reported engagement ranged from 3.01 to 3.42 and teacher-reported student engagement was 3.05 on 1- to 4-point scales, indicating high engagement in learning. Table 5 shows descriptive statistics based on factor scores.

Concordance and Discordance Between Measures of Engagement

Addressing Question 1, correlation coefficients between pairs of engagement variables showed (a) all correlations were positive,

Table 3
Confirmatory Factor Analysis for Teacher-Reported Behavioral Engagement

Item	Standardized factor loading
This student concentrates on doing his/her work during math class.	0.89
This student works as hard as he/she can during math class.	0.90
This student pays attention in math class.	0.89
This student tries to learn as much as he/she can about math.	0.89
This student's attention seems to wander during math class (reversed).	0.52
This student participates in discussions in math class.	0.72
This student asks off-topic questions during math class (reversed).	0.48
This student doesn't try very hard in math class (reversed).	0.79

Note. CFI = .95, TLI = .93, RMSEA = .04, SRMR = .04.

Table 4
Confirmatory Factor Analysis for Student-Reported Engagement

Item	Standardized factor loading
Cognitive engagement	
Today in math class I worked as hard as I could.	0.58
I thought about other things instead of math in math class today.	-0.33
Today I only paid attention in math when it was interesting.	-0.31
Today it was important to me that I understood the math really well.	0.68
I tried to learn as much as I could in math class today.	0.75
I did a lot of thinking in math class today.	0.64
Emotional engagement	
Math class was fun today.	0.80
Today I felt bored in math class.	-0.63
I enjoyed thinking about math today.	0.81
Learning math was interesting to me today.	0.82
I liked the feeling of solving problems in math today.	0.70
Social engagement	
Today I talked about math to other kids in class.	0.59
Today I helped other kids with math when they didn't know what to do.	0.72
Today I shared ideas and materials with other kids in math class.	0.65
Students in my math class helped each other learn today.	0.59

Note. CFI = .96, TLI = .96, RMSEA = .03, SRMR = .04.

statistically significant, and ranged from .11 to .68; (b) correlation coefficients were higher within each measure (range from .49 to .68) than between measures (range from .08 to .24); (c) lowest correlation values were between student-reported and teacher-reported engagement values (range from .11 to .24); and (d) consistent with the CFA results, correlation values confirm that students' view of their cognitive, emotional, and social engagement are related ($r = .49$ to $.68$) but have distinct characteristics (see Table 5).

Main Effect of Teacher–Student Interaction Quality and Student Gender on Engagement

Question 2 examined the extent to which quality of teacher–student interactions and student gender contributed to engagement. As a preliminary step, descriptive statistics for covariates and teacher–student interaction quality, as well as their correlation with engagement, were computed (see Table 6). Mean levels of emotional support and classroom organization appeared higher than those for instructional support. Teachers showed a slightly more limited range of emotional support and classroom organization compared to instructional support.

To address Question 2, we used the basic four-level model (labeled as Model 1 in Tables 7 and 8). We added one teacher-level variable (concurrent emotional support [Model 2A], classroom organization [Model 2B], or instructional support [Model 2C]) to the basic model. Table 7 shows results of the multilevel models for

Table 5
Intercorrelations and Descriptive Statistics for Engagement Variables

Variable	1	2	3	4	5
1. Observed behavioral engagement	—				
2. Teacher-reported behavioral engagement	.23** (356)	—			
3. Student-reported cognitive engagement	.17** (384)	.21** (356)	—		
4. Student-reported emotional engagement	.16** (384)	.11* (356)	.68** (384)	—	
5. Student-reported social engagement	.18** (384)	.24** (356)	.57** (384)	.49** (384)	—
<i>M</i>	0.00	3.05	3.42	3.30	3.01
<i>SD</i>	0.83	0.64	0.43	0.62	0.60
Min	-0.94	1.00	1.50	1.30	1.38
Max	3.04	4.00	4.00	4.00	4.00
<i>N</i>	384	359	384	384	384

Note. Sample sizes appear in parentheses.

* $p < .05$. ** $p < .01$.

observed and teacher-reported behavioral engagement. Students in classrooms with higher levels of classroom organization appeared more behaviorally engaged ($b = .13, p < .01$) than students in classrooms with lower levels of organization. Girls were observed to be more behaviorally engaged than boys ($b = .19, p < .01$). Pertaining to teacher-reported behavioral engagement, the quality of teacher-student interactions did not relate to teachers' report of students' engaged behavior. Teachers rated girls as more engaged than boys ($b = .11, p < .10$).

Table 8 shows results of the multilevel models for student-reported engagement. A consistent pattern of findings emerged: Students in classrooms with teachers who provided more emotional support and higher quality classroom organization reported higher cognitive engagement ($b = .03, p < .01; b = .03, p < .01$, respectively), higher emotional engagement ($b = .03, p < .01, b = .03, p < .01$, respectively), and higher social engagement ($b = .03, p < .01, b = .03, p < .01$, respectively). Girls reported higher cognitive and social engagement than boys ($b = .07, p < .01, b = .14, p < .01$, respectively).

Statistical Interactions Between Gender and Teacher-Student Interaction Quality

Question 3 queried the statistical interaction between quality of teacher-student interactions and gender in predicting engagement. The multilevel models included the basic model; the main effects (tested in Question 2, i.e., concurrent emotional support [Model 2A], classroom organization [Model 2B], or instructional support [Model 2C]); and one of three statistical interactions (Gender \times Emotional support [Model 3A], Gender \times Classroom organization [Model 3B], or Gender \times Instructional support [Model 3C]). As shown in Table 7, none of the interactions between gender and CLASS domains were statistically significant for observed or teacher-reported behavioral engagement. However, one Gender \times Teacher-student interaction quality effect emerged for each of the student-reported engagement outcomes, as shown in Models 3A, 3B, and 3C in Table 8. Analyses showed a small interaction effect between gender and classroom organization for student-reported cognitive engagement ($b = -0.06, p < .01$). As classroom orga-

Table 6
Correlations and Descriptive Statistics of Covariates With Engagement

Engagement	Child gender	Child age	Child FRPL	Initial achievement	Self-efficacy	Master's degree	Years exp.	CLASS ES	CLASS CO	CLASS IS
Observed factor score (behavioral)	.21**	-.06	.02	.20**	.08	.03	.06*	.16*	.26**	.08
Teacher-reported (behavioral)	.16**	-.08	.03	.29**	.18**	-.03	.01	.07	.11†	.03
Student-reported (cognitive)	.14*	-.06	.06	.11†	.32**	.06**	.15**	.10†	.04	.00
Student-reported (emotional)	.05	-.02	.13*	.03	.24**	-.03	.01	.05	.00	.02
Student-reported (social)	.10†	.07	.04	.18**	.40**	.06*	-.11**	.09	.03	-.05
<i>M</i>	0.53	10.47	0.22	506.38	3.30	0.63	12.34	5.16	5.99	3.37
<i>SD</i>	0.50	0.38	0.41	70.68	0.57	0.47	8.54	0.55	0.41	0.62
Min	0.00	8.24	0.00	284.00	1.40	0.00	1.00	3.83	4.39	1.83
Max	1.00	11.77	1.00	600.00	4.00	1.00	35.00	6.67	6.67	5.17
<i>N</i>	386	315	386	332	379	59	59	59	59	59

Note. Student gender (0 = male, 1 = female); FRPL (0 = no, 1 = yes). ES = Emotional support; CO = Classroom organization; IS = Instructional support. Sample size for correlations ranged from 300 to 382.

† $p < .10$. * $p < .05$. ** $p < .01$.

Table 7
Multilevel Model Results for Observed and Teacher-Reported Behavioral Engagement

Measure	Observed			Teacher-reported		
	b	SE	β	b	SE	β
Observation level (Model 1)						
Weeks of the year	0.00	0.01	0.01			
Weeks of the year ²	0.00	0.00	0.00			
Student level						
Child gender (female)	0.19**	0.05	0.22	0.11 [†]	0.06	0.17
Child age	-0.08	0.07	-0.03	-0.03	0.09	-0.02
Free/Reduced Price Lunch	0.07	0.06	0.08	0.14 [†]	0.07	0.17
Initial achievement (Math)	0.00**	0.00	0.14	0.00**	0.00	0.24
Self-efficacy	0.02	0.04	0.01	0.10*	0.05	0.09
Teacher level (Models 1, 2A, 2B, 2C)						
Master's degree (Model 1)	0.06	0.06	0.10	-0.04	0.08	-0.03
Years of experience (Model 1)	0.00	0.00	0.08	-0.00	0.00	-0.03
Concurrent emotional support (2A)	0.02	0.03	0.02	0.04	0.06	0.05
Concurrent classroom organization (2B)	0.13**	0.04	0.06	0.09	0.10	0.06
Concurrent instructional support (2C)	-0.01	0.02	0.00	0.02	0.06	0.03

Note. Teacher-reported behavioral engagement was collected only once and thus the model does not include the observation level. Results of analyses addressing Research Question 3 showed no significant interactions and therefore the interaction term is not reported.

[†] $p < .10$. * $p < .05$. ** $p < .01$.

nization increased, boys reported higher levels of cognitive engagement, but there was no comparable association evident for girls. Likewise, findings showed a statistically significant interaction effect between classroom organization and gender for student-reported emotional engagement ($b = -0.13$, $p < .01$). As classroom organization increased, boys reported higher emotional engagement, but girls did not. Post hoc analyses were conducted for both interaction effects. The slope of classroom organization

was statistically significant for boys ($p < .01$) but not girls ($p > .05$) for both outcomes. There was an interaction effect between instructional support and gender for student-reported social engagement ($b = -0.06$, $p < .05$). As instructional support increased, social engagement decreased for girls but not for boys. Post hoc analyses revealed a significant negative slope for girls ($p < .01$) but a nonsignificant slope ($p > .05$) for boys. (See Figure 1 for a description of these interactions.)

Table 8
Multilevel Model Results for Student-Reported Cognitive, Emotional, and Social Engagement

Measure	Cognitive			Emotional			Social		
	b	SE	β	b	SE	β	b	SE	β
Observation level (Model 1)									
Weeks of the year	-0.01**	0.00	-0.02	-0.02**	0.00	-0.03	-0.01	0.00	-0.01
Weeks of the year ²	0.00**	0.00	0.00	0.00**	0.00	0.00	-0.00*	0.00	-0.02
Student level									
Child gender (female)	0.07**	0.03	0.17	0.09	0.05	0.24	0.14**	0.04	0.24
Child age	-0.03	0.04	-0.02	-0.02	0.08	-0.01	0.04	0.06	0.03
Free/Reduced Price Lunch	0.06	0.04	0.14	0.17**	0.07	0.05	0.04	0.06	0.06
Initial achievement (math)	0.00	0.00	0.02	0.00	0.00	0.01	0.00	0.00	0.03
Self-efficacy	0.17**	0.03	0.23	0.26**	0.05	0.24	0.30**	0.04	0.28
Teacher level (Models 1, 2A, 2B, 2C)									
Master's degree (Model 1)	0.07**	0.03	0.20	0.05	0.05	0.13	0.08 [†]	0.05	0.13
Years of experience (Model 1)	-0.00	0.00	-0.06	-0.00	0.00	-0.06	-0.01**	0.00	-0.22
Concurrent emotional support (2A)	0.03**	0.01	0.04	0.03**	0.02	0.03	0.03**	0.01	0.02
Concurrent classroom organization (2B)	0.03**	0.01	0.03	0.03**	0.01	0.02	0.03**	0.01	0.02
Concurrent instructional support (2C)	-0.01	0.02	-0.01	-0.00	0.01	-0.01	-0.01	0.01	-0.01
Interactions (Models 3A, 3B, 3C)									
Gender × Emotional Support (3A)	—	—	—	—	—	—	—	—	—
Gender × Classroom Organization (3B)	-0.06**	0.02	-0.08	-0.13**	0.04	-0.09	—	—	—
Gender × Instructional Support (3C)	—	—	—	—	—	—	-0.06**	0.02	-0.06

Note. Student-reported engagement measures were collected in the fall, winter, and spring. Children = 387, teachers = 63, schools = 20. Interactions that were not statistically significant were not included in the final models and are not shown.

[†] $p < .10$. * $p < .05$. ** $p < .01$.

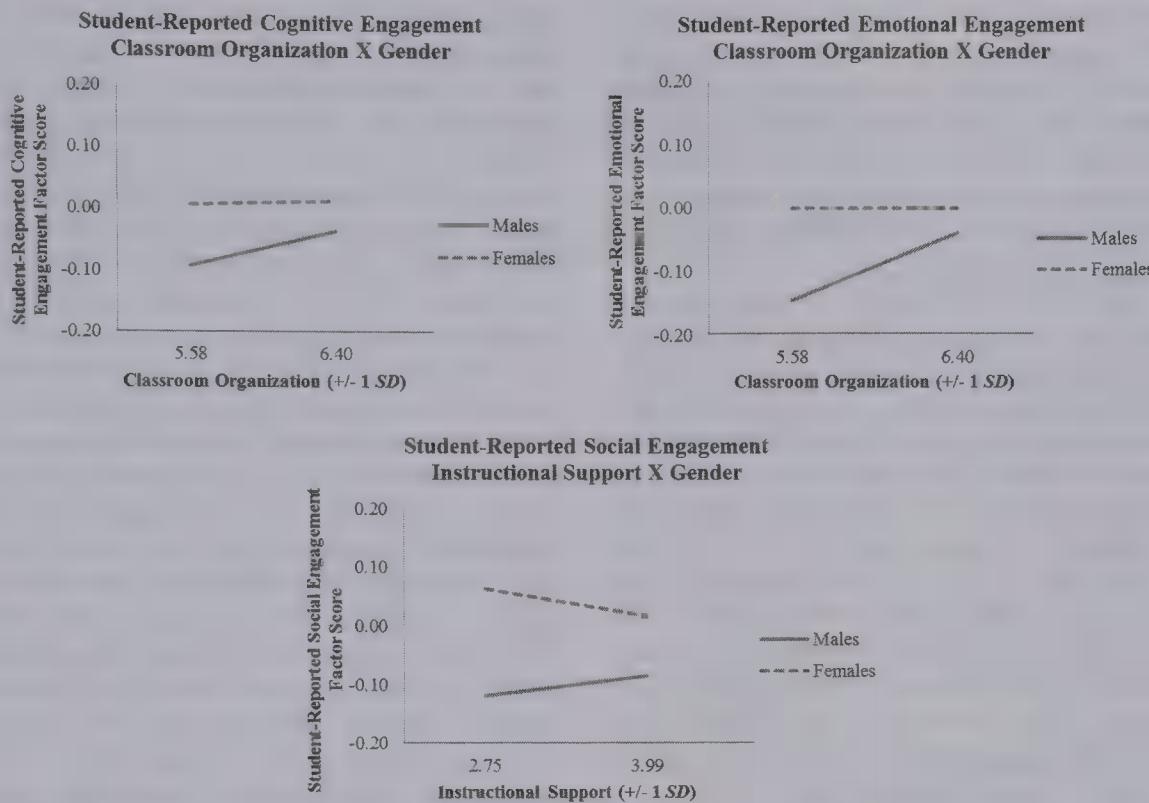


Figure 1. Interactions between teacher-student interaction quality (classroom organization and instructional support) and gender predicting student-reported engagement (cognitive, emotional, and social engagement).

Changes in Engagement Over Time

Although not the study's focus, results showed that the linear and curvilinear trends of time (weeks in the year) were not significant for observed behavioral engagement. Linear and curvilinear trends for time were present for all student-reported outcomes (cognitive, emotional, social engagement). For each student-reported outcome, the linear trend of time was significant (b from $-.01$ to $-.02$, $p < .01$) with a slight curvilinear relation ($b < .01$, $p < .01$), indicating that engagement decreased over time, but the decrease was steeper between Time 1 and 2 than 2 and 3.

Discussion

Three main findings emerged. First, the fifth graders, on average, showed high levels of math engagement, regardless of informant. Correlations between informants were lower than anticipated given the simultaneity of the data collection, a finding that suggests the unique vantage point of each informant. Second, the most systematic finding from the multilevel models was the link from teacher-student interaction quality to student-reported engagement. That is, students in classrooms with teachers who show warmth, caring, and individual responsiveness to their students reported working hard, enjoying learning about math, and sharing ideas and materials with other students in their classroom. Similarly, students in classroom with teachers who used proactive approaches to behavior management, facilitated smooth transitions between activities, and made learning objectives clear prior to learning also reported feeling greater cognitive, emotional, and social engagement in their math learning. Third, results showed higher engagement for girls than boys on three of the five engagement measures. Boys' report of their cognitive and emotional engagement was more closely coupled to the classroom con-

ditions (emotional and organizational support) than girls. An unexpected finding was that boys reported higher social engagement but girls reported lower social engagement in the presence of higher instructional support.

Measurement Concordance and Discordance

Correlation coefficients between different informants of student engagement were statistically significant, but small (1% to 11% shared variance). In contrast, associations within informants were high (24% to 49% of shared variance), even when comparing the same informant on different subconstructs of engagement. Findings match other literature showing modest cross-informant agreement (Gresham, Elliott, Cook, Vance, & Kettler, 2010; Konold & Pianta, 2007; Renk & Phares, 2004). Comparisons of informants can be considered in light of the integrative framework of motivation (Skinner et al., 2009). In theory, contexts, self-systems, and action are conceptually distinct, but in practice, accurate measurement of action (engagement) is challenging because it is tinged by characteristics of students' self-systems (goals, expectancies, perceived task value) and contexts (classroom interactions) depending on informant. The results provide researchers with new understanding as they consider measurement trade-offs.

We posit that low correlations among informants represent disparities in perspectives on the classroom and cannot be dismissed as error. For example, correlations between the observer's perception of behavioral engagement and students' feelings of engagement were low ($r = .24$ to $.26$), although the data were collected simultaneously. Behavioral engagement can be observed reliably by a research assistant and therefore may provide a more objective standpoint for understanding engagement. However, observed behavioral engagement may reflect superficial signs of engagement, whereas cognitive and

emotional engagement assessed by student ratings may reflect intrapsychic processes that, in part, are influenced by students' self-systems. This difference is important by the time students reach fifth grade because children have become accustomed to the student "script" and may show signs of behavioral engagement without genuine feelings of connection to learning. Researchers evaluating math curricula and interventions may benefit from gathering information about students' perception of engagement.

The majority of engagement research relies on teacher-reported behavioral engagement at the end of the year. Despite research linking teacher-reported behavioral engagement to achievement (e.g., Hughes & Kwok, 2007; Valiente, Lemery-Chalfant, Swanson, & Reiser, 2008), the present findings suggest that researchers should be cautious about overreliance on teacher-reported data. The correlations between observed behavioral engagement at three time points and teacher-reported behavioral engagement at the end of the year were low (from .29 to .33). Teacher-reported engagement is multidetermined and reflects students' actual engagement as well as teachers' attributes and perceptions (Mashburn et al. 2006). Teachers' rating tendencies may be systematic; for instance, teachers' ratings of fifth graders show greater inflation of girls' scores compared to boys (Robinson & Lubienski, 2011), and teachers appear to be better reporters of externalizing than internalizing problems (Konold & Pianta, 2007).

Contribution of Gender and Interaction Quality on Engagement

Girls were more engaged than boys for three of the five measured engagement constructs: observed behavioral engagement, student-reported cognitive engagement, and student-reported social engagement. Gender differences on self-reported emotional engagement and teacher-reported behavioral engagement approached statistical significance. Girls' higher observed behavioral engagement is consistent with other research suggesting higher behavioral engagement among girls than boys in late elementary and middle school (Marks, 2000; Wang, Willett, & Eccles, 2011). By definition, behavioral engagement involves the absence of disruptive behavior (Finn, Pannozzo, & Voelkl, 1995; Wang et al., 2011). The gender difference in observed behavioral engagement fits with other work describing more disruptive behavior in boys than girls (Finn et al., 1995). Girls reported higher cognitive engagement in math, comparable to findings in seventh graders (Wang et al., 2011).

The presence of emotional support was linked to students' own perception of their engagement (cognitive, emotional, and social). By definition, emotionally supportive teachers show warm and responsive behavior toward students and facilitate a classroom climate in which students exhibit positive, prosocial behavior (Pianta et al., 2008). Emotional support signals a sense of security to students that permits full attention to the academic work. It also fosters a classroom environment with positive communication and respect among peers (Luckner & Pianta, 2011). Both factors may be important for fifth graders facing challenging math learning. Findings match research pointing to the importance of the affective qualities of school, positive classroom climate, and teacher-student relationship for promoting engagement and learning (Borman & Overman, 2004; Decker, Dona, & Christenson, 2007; Dotterer & Lowe, 2011; Roorda et al., 2011; Stronge, Ward, & Grant, 2011; Reyes et al.,

2012). The finding that teachers' facilitation of a warm and supportive environment relates to students' perceived engagement but not higher observed or teacher-reported engagement underscores the point that student-reported engagement taps intrapsychic processes. The result also emphasizes the importance of emotionally supportive interactions between teachers and students in fifth grade, an issue that is crucial to convey to late elementary school math educators who are pressed for time or may perceive that relationship-building efforts are less essential for older students.

Well-organized classrooms appear to support students' engagement in math learning, as evidenced by higher observed behavioral engagement and student-reported cognitive and emotional engagement. The finding linking classroom organization to observed behavioral engagement is not surprising; interesting learning formats, clear statement of expectations, high productivity are teaching practices that have been linked to observed student behavioral engagement in classic (Brophy, 1983) and recent work (Downer, Rimm-Kaufman, & Pianta, 2007). However, the result that higher classroom organization relates to students' perception of their cognitive engagement (commitment to paying attention, desire to understand complicated material) and emotional engagement (enjoyment of math class and problem solving) stands out as important new contribution. By fifth grade, teachers may perceive students' need for more autonomy. Effective practices attuned to fifth graders' developmental needs involve fostering autonomy while maintaining clear objectives and minimizing classroom chaos (Eccles, 2004).

Counter to expectation, results showed no main effects of instructional support on students' engagement in learning. This is a surprising finding. One possible explanation stems from the reliance on the CLASS, a global measure of instructional support that reflects teachers' interactions with their whole classroom of students. Students within a single classroom show a wide range of abilities. Although teachers may be providing even amounts of concept development or high quality feedback to students across the classroom, the level of instruction may be too hard for some students, too easy for others, and just right for others. Another explanation pertains to the reliance on an observational measure of instructional support. Gathering information on each student's perception of instructional support from their teacher may increase accuracy. Future work is needed that considers students' ability level relative to the level of the mathematical tasks and taps students' perception of teachers' instructional support.

None of the three domains of teacher-student interaction quality related to teacher-reported behavioral engagement. Measuring teacher-student interactions and student engagement concurrently reveals associations between teacher and student behavior that otherwise may be masked in an end-of-the-year, teacher-reported measure. Teachers' report of engagement may reflect teacher attributes as well as student engagement (Mashburn et al., 2006).

Interactions Between Teacher-Student Interaction Quality and Gender

Classroom organization was associated with students' perception of their engagement more for boys than girls. Boys may be

more distracted by chaotic learning environments and thus show more difficulty engaging in learning (Ponitz, Rimm-Kaufman, Brock, & Nathanson, 2009). Girls may have better developed self-regulatory skills and require fewer external structures to support their engagement.

Instructional support findings were surprising. Girls in classrooms with higher instructional support reported lower social engagement, whereas there was no relation between instructional support and social engagement in boys. The finding may reflect complementarity between teacher-student and student-student interactions. High levels of instructional support may be more evident in classrooms where teachers engage in frequent, high quality interactions with students but facilitate fewer peer interactions. In fact, two of the three instructional support CLASS dimensions (quality of feedback, language modeling) involve verbal interactions between teachers and students. Frequent teacher-student interactions may supplant peer-to-peer interactions that would occur otherwise (Patrick et al., 2007). It is unclear why this association was present among girls but not boys.

Limitations

Several limitations require mention. First, we did not emphasize peer interactions. By fifth grade, students are increasingly aware and influenced by peers. Peer liking and acceptance contribute to teacher-student interaction quality, peer nominations of academic competence, and students' perception of their achievement self-efficacy (Hughes & Chen, 2011). Fifth graders may be sensitive to classroom composition. The engagement of a student's peer group links to their engagement over the course of the year, net of other factors—a consideration not included in the present work (Kindermann, 2007). Second, data were gathered in the context of a descriptive study; therefore, the work does not support causal inferences. Future research using an experimental design is needed. Third, the three approaches to measuring engagement reflect different time sampling. Observational data were based on 24 min spread across 3 days; student-reported data were based on students' reflection on the full period of math class across those same 3 days. Teacher-reported measures were based on reflections of students over the course of the year. Fourth, the data collection did not include students' reports of their teachers' quality of interactions. Fifth, some of the measures have lower than ideal reliability.

Closing Comments

Recommendations for improving mathematics achievement hinge on teachers' ability to engage students in learning in the classroom (CCSSI, 2014; NCTM, 2000; National Research Council, 2005), raising questions about the extent to which different types of teacher-student interactions contribute to enhanced engagement in the math classroom. On a daily basis, teachers rely on their perception of students to know whether to adjust the content and pace of learning to keep students engaged. However, by fifth grade students know how to appear interested and engaged, leaving teachers with questions about what they can do to be sure that students are putting forth their best effort and are truly curious and interested in the math. Findings lead to at least two implications. Teachers having difficulty gauging students' interest, curiosity and

attention toward math may want to rely less on their own insights or observations of an observer and generate strategies for receiving direct and honest feedback from their students. Despite the fact that fifth graders are not young children, the students, especially boys, appear to be well attuned to the warmth and responsiveness of their teacher and clarity of expectations in the classroom.

References

Avant, T. S., Gazelle, H., & Faldowski, R. (2011). Classroom emotional climate as a moderator of anxious solitary children's longitudinal risk for peer exclusion: A Child \times Environment model. *Developmental Psychology, 47*, 1711–1727. doi:10.1037/a0024021

Bodovski, K., & Farkas, G. (2007). Mathematics growth in early elementary school: The roles of beginning knowledge, student engagement, and instruction. *The Elementary School Journal, 108*, 115–130. doi:10.1086/525550

Bohn, C. M., Roehrig, A. D., & Pressley, M. (2004). The first days of school in the classrooms of two more effective and four less effective primary-grades teachers. *The Elementary School Journal, 104*, 269–287. doi:10.1086/499753

Borman, G. D., & Overman, L. T. (2004). Academic resilience in mathematics among poor and minority students. *The Elementary School Journal, 104*, 177–195. doi:10.1086/499748

Bronson, M. B. (2001). *Self-regulation in early childhood: Nature and nurture*. New York, NY: Guilford Press.

Brophy, J. (1983). Classroom organization and management. *The Elementary School Journal, 83*, 265–286. doi:10.1086/461318

Brophy, J., & Good, T. (1986). Teacher behavior and student achievement. In M. C. Whitrock (Ed.), *The handbook of research on teaching* (3rd ed., pp. 328–375). New York, NY: Macmillan.

Cameron, C. E., Connor, C. M. D., & Morrison, F. J. (2005). Effects of variation in teacher organization on classroom functioning. *Journal of School Psychology, 43*, 61–85. doi:10.1016/j.jsp.2004.12.002

Christenson, S. L., Reschly, A. L., & Wylie, C. (2012). *Handbook of research on student engagement*. New York, NY: Springer.

Common Core State Standards Initiative. (2014). *Mathematics standards*. Retrieved from <http://www.corestandards.org/math>

Connell, J. P., & Wellborn, J. G. (1991). Competence, autonomy, and relatedness: A motivational analysis of self-system processes. In M. R. Gunnar & L. A. Sroufe (Eds.), *Self processes and development: The Minnesota Symposia on Child Psychology* (Vol. 23, pp. 43–77). Hillsdale, NJ: Erlbaum.

Crosnoe, R., Morrison, F., Burchinal, M., Pianta, R., Keating, D., Friedman, S. L., & Clarke-Stewart, K. A. (2010). Instruction, teacher-student relations, and math achievement trajectories in elementary school. *Journal of Educational Psychology, 102*, 407–417. doi:10.1037/a0017762

Curby, T. W., Rimm-Kaufman, S. E., & Abry, T. (2013). Do emotional support and classroom organization earlier in the year set the stage for higher quality instruction? *Journal of School Psychology, 51*, 557–569. doi:10.1016/j.jsp.2013.06.001

Decker, D. M., Dena, D. P., & Christenson, S. L. (2007). Behaviorally at-risk African American students: The importance of student-teacher relationships for student outcomes. *Journal of School Psychology, 45*, 83–109. doi:10.1016/j.jsp.2006.09.004

Dotterer, A. M., & Lowe, K. (2011). Classroom context, school engagement, and academic achievement in early adolescence. *Journal of Youth and Adolescence, 40*, 1649–1660. doi:10.1007/s10964-011-9647-5

Downer, J. T., Rimm-Kaufman, S. E., & Pianta, R. C. (2007). How do classroom conditions and children's risk for school problems contribute to children's behavioral engagement in learning? *School Psychology Review, 36*, 413–432.

Eccles, J. S. (2004). Schools, academic motivation, and stage-environment

fit. In R. M. Lerner & L. D. Steinberg (Eds.), *Handbook of adolescent psychology* (pp. 125–153). Hoboken, NJ: Wiley.

Evans, G. W., & English, K. (2002). The environment of poverty: Multiple stressor exposure, psychophysiological stress, and socioemotional adjustment. *Child Development*, 73, 1238–1248. doi:10.1111/1467-8624.00469

Evans, G. W., & Rosenbaum, J. (2008). Self-regulation and the income-achievement gap. *Early Childhood Research Quarterly*, 23, 504–514. doi:10.1016/j.ecresq.2008.07.002

Finn, J. D., Pannozzo, G. M., & Voelkl, K. E. (1995). Disruptive and inattentive-withdrawn behavior and achievement among fourth graders. *The Elementary School Journal*, 95, 421–434. doi:10.1086/461853

Finn, J. D., & Zimmer, K. S. (2012). Student engagement: What is it? Why does it matter? In S. L. Christenson, A. L. Reschly, & C. Wylie (Eds.), *Handbook of research on student engagement* (pp. 97–131). New York, NY: Springer. doi:10.1007/978-1-4614-2018-7_5

Fredricks, J. A., Blumenfeld, P. C., & Paris, A. H. (2004). School engagement: Potential of the concept, state of the evidence. *Review of Educational Research*, 74, 59–109. doi:10.3102/00346543074001059

Furrer, C., & Skinner, E. (2003). Sense of relatedness as a factor in children's academic engagement and performance. *Journal of Educational Psychology*, 95, 148–162. doi:10.1037/0022-0663.95.1.148

Fuson, K., Kalchman, M., & Bransford, J. (2005). Mathematical understanding: An introduction. In M. S. Donovan & J. D. Bransford (Eds.), *How students learn: History, mathematics, and science in the classroom* (pp. 217–256). Washington, DC: National Academies Press.

Gest, S. D., Domitrovich, C. E., & Welsh, J. A. (2005). Peer academic reputation in elementary school: Associations with changes in self-concept and academic skills. *Journal of Educational Psychology*, 97, 337–346. doi:10.1037/0022-0663.97.3.337

Greenwood, C. R., Horton, B. T., & Utley, C. A. (2002). Academic engagement: Current perspectives on research and practice. *School Psychology Review*, 31, 328–349.

Gresham, F. M., Elliott, S. N., Cook, C. R., Vance, M. J., & Kettler, R. (2010). Cross-informant agreement for ratings for social skill and problem behavior ratings: An investigation of the Social Skills Improvement System—Rating Scales. *Psychological Assessment*, 22, 157–166. doi:10.1037/a0018124

Henningsen, M., & Stein, M. K. (1997). Mathematical tasks and student cognition: Classroom-based factors that support and inhibit high-level mathematical thinking and reasoning. *Journal for Research in Mathematics Education*, 28, 524–549. doi:10.2307/749690

Hiebert, J., & Grouws, D. (2007). The effects of classroom mathematics teaching on students' learning. In F. K. Lester (Ed.), *Second handbook of research on mathematics teaching and learning* (Vol. 1, pp. 371–404). Charlotte, NC: Information Age.

Hughes, J., & Chen, Q. (2011). Reciprocal effects of student-teacher and student-peer relatedness: Effects on academic self-efficacy. *Journal of Applied Developmental Psychology*, 32(5), 278–298.

Hughes, J., & Kwok, O. (2007). Influence of student-teacher and parent-teacher relationships on lower achieving readers' engagement and achievement in the primary grades. *Journal of Educational Psychology*, 99, 39–51. doi:10.1037/0022-0663.99.1.39

Kindermann, T. A. (2007). Effects of naturally existing peer groups on changes in academic engagement in a cohort of sixth graders. *Child Development*, 78, 1186–1203. doi:10.1111/j.1467-8624.2007.01060.x

Kline, R. B. (2011). *Principles and practice of structural equation modeling* (3rd ed.). New York, NY: Guilford Press.

Kong, Q. P., Wong, N. Y., & Lam, C. C. (2003). Student engagement in mathematics: Development of instrument and validation of construct. *Mathematics Education Research Journal*, 15, 4–21. doi:10.1007/BF03217366

Konold, T. R., & Pianta, R. C. (2007). The influence of informants on ratings of children's behavioral functioning. *Journal of Psychoeducational Assessment*, 25, 222–236. doi:10.1177/0734282906297784

Ladd, G. W., Birch, S. H., & Buhs, E. S. (1999). Children's social and scholastic lives in kindergarten: Related spheres of influence? *Child Development*, 70, 1373–1400. doi:10.1111/1467-8624.00101

Linnenbrink, E. A., & Pintrich, P. R. (2003). The role of self-efficacy beliefs in student engagement and learning in the classroom. *Reading & Writing Quarterly*, 19, 119–137. doi:10.1080/10573560308223

Little, R., & Rubin, D. (1987). *Statistical analysis with missing data* (2nd ed.). New York, NY: Wiley.

Luckner, A. E., & Pianta, R. C. (2011). Teacher-student interactions in fifth grade classrooms: Relations with children's peer behavior. *Journal of Applied Developmental Psychology*, 32, 257–266. doi:10.1016/j.appdev.2011.02.010

Marks, H. M. (2000). Student engagement in instructional activity: Patterns in the elementary, middle, and high school years. *American Educational Research Journal*, 37, 153–184. doi:10.3102/00028312037001153

Martin, A. J., Anderson, J., Bobis, J., Way, J., & Vellar, R. (2012). Switching on and switching off in mathematics: An ecological study of future intent and disengagement among middle school students. *Journal of Educational Psychology*, 104, 1–18. doi:10.1037/a0025988

Mashburn, A. J., Hamre, B. K., Downer, J. T., & Pianta, R. C. (2006). Teacher and classroom characteristics associated with teachers' ratings of prekindergartners' relationships and behaviors. *Journal of Psychoeducational Assessment*, 24, 367–380. doi:10.1177/0734282906290594

Matsumura, L. C., Slater, S. C., & Crosson, A. (2008). Classroom climate, rigorous instruction and curriculum, and students' interactions in urban middle schools. *The Elementary School Journal*, 108, 293–312. doi:10.1086/528973

Meece, J. (2009). *Measure of Student Cognitive Engagement*. Unpublished measure, University of North Carolina.

Midgley, C., Maehr, M. L., Hruda, L. Z., Anderman, E., Anderman, L., Freeman, K. E., . . . Urdan, T. (2000). *Manual for the Patterns of Adaptive Learning Scale*. Ann Arbor: University of Michigan.

Muthén, L. K., & Muthén, B. O. (2010). *Mplus user's guide* (6th ed.). Los Angeles, CA: Muthén & Muthén.

National Council of Teachers of Mathematics. (2000). *Principles and standards for school mathematics* (Vol. 1). Ann Arbor, MI: Author.

National Research Council. (2001). *Adding it up: Helping children learn mathematics*. Washington, DC: The National Academies Press.

National Research Council. (2005). *How students learn: History, mathematics, and science in the classroom*. Washington, DC: The National Academies Press.

NICHD Early Child Care Research Network. (2005). A day in third grade: Classroom quality, teacher, and student behaviors. *The Elementary School Journal*, 105, 305–323. doi:10.1086/428746

Patrick, H., Ryan, A. M., & Kaplan, A. (2007). Early adolescents' perceptions of the classroom social environment, motivational beliefs, and engagement. *Journal of Educational Psychology*, 99, 83–98. doi:10.1037/0022-0663.99.1.83

Pianta, R. C., & Hamre, B. K. (2009). Conceptualization, measurement, and improvement of classroom processes: Standardized observation can leverage capacity. *Educational Researcher*, 38, 109–119. doi:10.3102/0013189X09332374

Pianta, R. C., La Paro, K. M., & Hamre, B. K. (2008). *Classroom Assessment Scoring System (CLASS: PreK-3)*. Baltimore, MD: Brookes.

Ponitz, C. C., Rimm-Kaufman, S. E., Brock, L. L., & Nathanson, L. (2009). Early adjustment, gender differences, and classroom organizational climate in first grade. *The Elementary School Journal*, 110, 142–162. doi:10.1086/605470

Ponitz, C. C., Rimm-Kaufman, S. E., Grimm, K. J., & Curby, T. W. (2009). Kindergarten classroom quality, behavioral engagement, and reading achievement. *School Psychology Review*, 38, 102–120.

Raphael, L. M., Pressley, M., & Mohan, L. (2008). Engaging instruction in middle school classrooms: An observational study of nine teachers. *The Elementary School Journal*, 109, 61–81. doi:10.1086/592367

Renk, K., & Phares, V. (2004). Cross-informant ratings of social competence in children and adolescents. *Clinical Psychology Review*, 24, 239–254. doi:10.1016/j.cpr.2004.01.004

Reschly, A., & Christenson, S. L. (2012). Jingle, jangle, and conceptual haziness: Evolution and future directions of the engagement construct. In S. L. Christenson, A. L. Reschly, & C. Wylie (Eds.), *Handbook of research on student engagement* (pp. 3–19). New York, NY: Springer.

Reyes, M. R., Brackett, M. A., Rivers, S. E., White, M., & Salovey, P. (2012). Classroom emotional climate, student engagement and academic achievement. *Journal of Educational Psychology*, 104, 700–712. doi:10.1037/a0027268

Rimm-Kaufman, S. E., Curby, T. W., Grimm, K. J., Nathanson, L., & Brock, L. L. (2009). The contribution of children's self-regulation and classroom quality to children's adaptive behaviors in the kindergarten classroom. *Developmental Psychology*, 45, 958–972. doi:10.1037/a0015861

Rimm-Kaufman, S. E., Early, D. M., Cox, M., Saluja, G., Pianta, R., Bradley, R., & Payne, C. (2002). Early behavioral attributes and teachers' sensitivity as predictors of competent behavior in the kindergarten classroom. *Journal of Applied Developmental Psychology*, 23, 451–470.

Rimm-Kaufman, S. E., & Hamre, B. K. (2010). The role of psychological and developmental science in efforts to improve teacher quality. *Teachers College Record*, 112, 2988–3023.

Robinson, J. P., & Lubienski, S. T. (2011). The development of gender achievement gaps in mathematics and reading during elementary and middle school. *American Educational Research Journal*, 48, 268–302. doi:10.3102/0002831210372249

Roorda, D. L., Koomen, H. M. Y., Spilt, J. L., & Oort, F. J. (2011). The influence of affective teacher–student relationships on students' school engagement and achievement. *Review of Educational Research*, 81, 493–529. doi:10.3102/0034654311421793

Rowley, S. J., Kurtz-Costes, B., Meyer, R., & Kizzie, K. (2009). *Engagement and self-concept during the transition to middle school: Gender and domain-specific differences in change in African American youth*. Unpublished manuscript, University of Michigan.

Rudasill, K., Gallagher, K. C., & White, J. M. (2010). Temperamental attention and activity, classroom emotional support, and academic achievement in third grade. *Journal of School Psychology*, 48, 113–134. doi:10.1016/j.jsp.2009.11.002

Schoenfeld, A. H. (1992). Learning to think mathematically: Problem solving, metacognition, and sense making in mathematics. In D. Grouws (Ed.), *Handbook of research on mathematics teaching and learning* (pp. 334–370). New York, NY: Macmillan.

Schunk, D., & Pajares, F. (2005). Competence perceptions and academic functioning. In A. J. Elliot (Ed.), *Handbook of competence and motivation* (pp. 85–104). New York, NY: Guilford Press.

Skinner, E. A., & Belmont, M. J. (1993). Motivation in the classroom: Reciprocal effects of teacher behavior and student engagement across the school year. *Journal of Educational Psychology*, 85, 571–581. doi:10.1037/0022-0663.85.4.571

Skinner, E., Furrer, C., Marchand, G., & Kindermann, T. (2008). Engagement and disaffection in the classroom: Part of a larger motivational dynamic? *Journal of Educational Psychology*, 100, 765–781. doi:10.1037/a0012840

Skinner, E. A., Kindermann, T. A., Connell, J. P., & Wellborn, J. G. (2009). Engagement and disaffection as organizational constructs in the dynamics of motivational development. In K. R. Wentzel & A. Wigfield (Eds.), *Handbook of motivation at school* (pp. 223–245). New York, NY: Routledge.

Stronge, J. H., Ward, T. J., & Grant, L. W. (2011). What makes good teachers good? A cross-case analysis of the connection between teacher effectiveness and student achievement. *Journal of Teacher Education*, 62, 339–355. doi:10.1177/0022487111404241

Tucker, C. M., Zayco, R. A., Herman, K. C., Reinke, W. M., Trujillo, M., Carraway, K., . . . Ivery, P. D. (2002). Teacher and child variables as predictors of academic engagement among low-income African American children. *Psychology in the Schools*, 39, 477–488. doi:10.1002/pits.10038

Valiente, C., Lemery-Chalfant, K., Swanson, J., & Reiser, M. (2008). Prediction of children's academic competence from their effortful control, relationships, and classroom participation. *Journal of Educational Psychology*, 100, 67–77. doi:10.1037/0022-0663.100.1.67

Virginia Department of Education. (2008). *Virginia Standards of Learning technical report: 2008–2009 administration cycle*. Retrieved from http://www.doe.virginia.gov/testing/test_administration/technical_reports/sol_technical_report_2008-09_administration_cycle.pdf

Virginia Department of Education. (2010). *Virginia standards of learning assessment: Test blueprint, Grade 4 mathematics*. Retrieved from http://www.doe.virginia.gov/testing/sol/blueprints/mathematics_blueprints/2009/blueprint_math4%20.pdf

Voelkl, K. E. (1995). School warmth, student participation, and achievement. *Journal of Experimental Education*, 63, 127–138. doi:10.1080/00220973.1995.9943817

Wang, M. T., Willett, J. B., & Eccles, J. S. (2011). The assessment of school engagement: Examining dimensionality and measurement invariance by gender and race/ethnicity. *Journal of School Psychology*, 49, 465–480. doi:10.1016/j.jsp.2011.04.001

Woodward, J., Beckmann, S., Driscoll, M., Franke, M., Herzig, P., Jitendra, A., . . . Ogbuehi, P. (2012). *Improving mathematical problem solving in Grades 4 through 8: A practice guide* (NCEE 2012–4055). Retrieved from http://ies.ed.gov/ncee/wwc/pdf/practice_guides/mps_pg_052212.pdf

Wu, J., Hughes, J., & Kwok, O. (2010). Teacher–student relationship quality type in elementary grades: Effects on trajectories for achievement and engagement. *Journal of School Psychology*, 48, 357–387. doi:10.1016/j.jsp.2010.06.004

Received December 22, 2012

Revision received May 2, 2014

Accepted May 12, 2014 ■

“Michael Can’t Read!” Teachers’ Gender Stereotypes and Boys’ Reading Self-Concept

Jan Retelsdorf

Leibniz Institute for Science and Mathematics Education,
Kiel, Germany

Katja Schwartz

University of Kiel

Frank Asbrock

Philipps University Marburg

According to expectancy-value theory, the gender stereotypes of significant others such as parents, peers, or teachers affect students’ competence beliefs, values, and achievement-related behavior. Stereotypically, gender beliefs about reading favor girls. The aim of this study was to investigate whether teachers’ gender stereotypes in relation to reading—their belief that girls outperform boys—have a negative effect on the reading self-concept of boys, but not girls. We drew on a longitudinal study comprising two occasions of data collection: toward the beginning of Grade 5 (T1) and in the second half of Grade 6 (T2). Our sample consisted of 54 teachers and 1,358 students. Using multilevel modeling, controlling for T1 reading self-concept, reading achievement, and school track, we found a negative association between teachers’ gender stereotype at T1 and boys’ reading self-concept at T2, as expected. For girls, this association did not yield a significant result. Thus, our results provide empirical support for the idea that gender differences in self-concept may be due to the stereotypical beliefs of teachers as significant others. In concluding, we discuss what teachers can do to counteract the effects of their own gender stereotypes.

Keywords: gender stereotypes, reading self-concept

Gender differences in students’ academic self-concepts often exceed differences in actual achievement (Hyde & Durik, 2005). Drawing on expectancy-value theory (e.g., Eccles & Wigfield, 2002; Wigfield & Eccles, 2000), one compelling explanation of this discrepancy is that self-concepts develop, *inter alia*, as a function of the gender beliefs or stereotypes of significant others such as parents, peers, or teachers. Stereotypes are very powerful in shaping biased expectations of and behaviors toward groups, especially in regard to broad categories like gender (Schneider, 2004). Such expectations and behaviors can in turn affect the self-concept of members of the stereotyped group. This is in line with the assumption of social identity theory (Tajfel & Turner, 1986) that widely held stereotypes about social groups can influ-

ence a person’s view of her- or himself. People derive their identity in part from the social group they belong to and therefore from socially shared beliefs about their group’s characteristics (cf. Tajfel, 1981). For example, girls may develop a positive verbal self-concept due in part to their knowledge of the social belief that girls and women are good at language-related tasks. Regarding educational outcomes and gender, the question as to which group is negatively stereotyped depends on the domain (Plante, de la Sablonnière, Aronson, & Théorêt, 2013). Whereas there has been some research on the negative effects of stereotyping for girls in mathematics (see e.g., Nguyen & Ryan, 2008, for a review), little is known about the negative effects of stereotypes for boys in reading. In this longitudinal study, we aimed to investigate the relation of teachers’ gender stereotypes about reading as a stereotypically female academic outcome (Schmenk, 2004) to students’ self-concept in reading. There has as yet not been much research testing the assumption of expectancy-value theory that teachers’ gender stereotypes may explain gender differences in students’ reading self-concept.

This article was published Online First June 16, 2014.

Jan Retelsdorf, Department of Educational Research, Leibniz Institute for Science and Mathematics Education, Kiel, Germany; Katja Schwartz, Department of Psychology, University of Kiel; Frank Asbrock, Department of Psychology, Philipps University Marburg.

The research reported in this article is part of the project “Self-Concept, Motivation, and Literacy: Development of Student Reading Behavior,” directed by Jens Möller (Christian-Albrechts-University of Kiel). The project was funded by the German Research Foundation (DFG; Mo 648/15-1/15-3). We would like to thank Stephen McLaren for his editorial support during preparation of this article.

Correspondence concerning this article should be addressed to Jan Retelsdorf, Leibniz Institute for Science and Mathematics Education, Olshausenstr. 62, D-24118 Kiel, Germany. E-mail: jretelsdorf@ipn.uni-kiel.de

Gender Differences in the Development of Language-Related Self-Concepts

Gender is believed to play an important role in shaping students’ ability self-concepts (Eccles & Wigfield, 2002; Meece, Bower Glienke, & Burg, 2006; Wigfield & Eccles, 2000). Since ability self-concepts are highly domain specific (Marsh, Trautwein, Lüdtke, Köller, & Baumert, 2006; Möller, Retelsdorf, Köller, & Marsh, 2011), the question as to which gender is advantaged and which disadvantaged depends of

course on the particular domain. Typically, ability self-concept is higher for the gender that is stereotypically favored in a particular domain (Watt & Eccles, 2008). Thus, boys are believed to have higher mathematics and related self-concepts, and girls to have higher language-related self-concepts. Indeed, there is compelling evidence that girls report higher confidence in their language abilities than do boys (Durik, Vida, & Eccles, 2006; Eccles, Wigfield, Harold, & Blumenfeld, 1993; Ireson & Hallam, 2009; Wigfield et al., 1997) although not all studies have found such differences (Anderman et al., 2001; Evans, Coping, Rowley, & Kurtz-Costes, 2011; Skaalvik & Skaalvik, 2004). Moreover, there is even some evidence from longitudinal studies that these gender differences increase over time. Jacobs, Lanza, Osgood, Eccles, and Wigfield (2002) reported such a widening gap between girls' and boys' language-related self-concept from Grades 1 to 12. In another longitudinal study, Archambault, Eccles, and Vida (2010) identified seven groups of children with distinct trajectories of language-related self-concept. They found a higher proportion of girls maintained the highest and most stable self-concepts over time; conversely, a higher proportion of boys indicated substantial self-concept decline. These results also indicated an increasing gender difference over time. It is also noteworthy, however, that self-concepts decline for both boys and girls over time. Thus, the widening gender gap would appear to be a result of the steeper decline within the group of boys.

A promising approach to the explanation of gender differences in self-concept is provided by Eccles' expectancy-value theory of achievement-related choices (e.g., Eccles et al., 1983; Eccles & Wigfield, 2002; Wigfield & Eccles, 2000). This theory, which provides a general model for the explanation of achievement-related choices and behaviors, has a particular focus on the understanding of gender differences. The model deals with the question of the circumstances under which a person will undertake a challenging achievement task. This is explained in terms of high value of the task and high expectation of success. Moreover, the model also provides a valuable framework for the explanation of gender differences in ability self-concepts that are closely related to one core variable of the model—expectation of success. According to expectancy-value theory, a person's self-concept is shaped not only by his or her previous achievement but also by a variety of social and cultural factors. These factors comprise cultural gender roles that prescribe certain behaviors as appropriate or inappropriate for males or females, as well as gender stereotypes. Moreover, the behaviors and beliefs of significant others, such as peers, parents, and teachers, play an important role in shaping students' self-concepts. In the present research, we focused on the role of teachers, as there is some evidence that teachers can contribute to the gender gap. For example, they may pay more attention to boys than to girls (DeZolt & Hull, 2001) and communicate overall more with boys than with girls—in particular, approving boys' academic behavior and disapproving their social behavior more frequently (Swinson & Harrop, 2009). However, there has been little research directly connecting teachers' gender beliefs with student outcomes. In the present research, we addressed this lacuna by investigating the effect of teachers' gender stereotypes about reading on students' self-concepts.

Gender Stereotypes in Education

Stereotypes can be broadly defined as "shared beliefs about personality traits and behaviors of group members" (Fiedler & Bless, 2001, p. 123). Stereotyping results from categorizing individuals into groups, according to their presumed common attributes. While stereotypes can function as cognitive schemas to facilitate social interactions with unknown individuals, as overgeneralizations of traits for a group in general, they also shape expectations and behaviors. Consensually shared stereotypes within a culture can serve as social norms for behavior toward the stereotyped group (e.g., Asbroek, Nieuwoudt, Duckitt, & Sibley, 2011; Cuddy, Fiske, & Glick, 2007). In respect to gender, the two groups, males and females, are presumed to differ in their traits, abilities, and motivation (cf. Schmenk, 2004). The latter two are of particular interest in education while, as mentioned earlier, stereotypes depend on the particular domain that is being considered. Research investigating gender stereotypes in the educational context has mainly focused on *stereotype threat*—a phenomenon describing how stereotypes can become self-fulfilling in a particular situation (Aronson & Steele, 2005; Steele, 1997). Stereotype threat means a situational threat due to a negative stereotype about one's own group (Steele, 1997). In the educational context, stereotyped persons feel extra pressure not to fail in a situation where academic competence is relevant. Regarding gender, there is quite strong evidence—mainly from experimental research—for the negative impact of stereotype threat on the performance of girls or women in mathematics tests (e.g., Nguyen & Ryan, 2008 for a review). Moreover, in a recent study, Hartley and Sutton (2013) investigated the role of stereotype threat in boys' general academic underachievement. In one study, they showed that girls and boys believed that girls academically outperform boys and also thought that adults believed this to be true. In a second study, they manipulated stereotype threat by telling the children in their sample that boys tend to perform lower than girls at school. This manipulation negatively affected the boys' performance in reading, writing, and mathematics but had no effect on girls' performance.

Moreover, Plante et al. (2013) have investigated students' own gender stereotypes and their associations with self-concept, task values, and achievement in a naturalistic setting. They tested the hypothesis from expectancy-value theory (e.g., Eccles & Wigfield, 2002) that the relationship between gender stereotypes and academic outcomes is mediated by students' self-concepts and task values in the corresponding domain. In their cross-sectional study, they found that effects of gender stereotypes on achievement in mathematics and language arts were mediated by students' self-concepts and task values. However, Plante et al. (2013) only investigated the students' own stereotypes. Thus, the idea of expectancy-value theory, that the gender beliefs of significant others affect students' self-concept development, could not be tested. Generally, this is an under-researched issue. Even though stereotypes have been a "hot topic" in general, as Jussim, Eccles, and Madon (1996) realized, only a few studies have investigated the effects of stereotypes of significant others in more naturalistic settings. To the best of our knowledge, there has not been much development in the research since this conclusion was drawn. One notable exception is research showing that parents' stereotypic beliefs affect children's perceptions of their ability. For example, Jacobs and Eccles (1992) have shown that across three domains—

mathematics, sports, and social domain—mothers' gender stereotypes either lead to an overestimated perception of their child's ability, if the child is stereotypically favored, or to an underestimation of their child's ability, if the child is stereotypically disadvantaged. In turn, the mothers' perceptions of their child's ability affect the children's own perception of their ability. Similarly, Tiedemann (2000) found that mothers' and fathers' gender stereotypes predicted their beliefs about their child's abilities, which in turn were related to their child's self-perceptions of ability. More recently, Rouland, Rowley, and Kurtz-Costes (2013) found that parents' gender stereotypes were related to their attributions for their children's academic successes and failures that in turn were related to the children's own self-beliefs.

However, less is known about the effects of stereotypes in other groups of significant others. In the educational context, of course, one of the most important groups is that of teachers, because they interact with children on a daily basis, instruct them, judge them, and—as a consequence—develop evaluations of the children's cognitive and social development and long-term career prospects. Indeed, there has been a vast amount of research on the related issue of teacher expectations for low- and high-achieving students (for a review, see Jussim & Harber, 2005). While there has been some research on student gender as a potential moderator of teacher expectation effects (e.g., de Boer, Bosker, & van der Werf, 2010), there are fewer studies on teachers' explicit beliefs about boys' and girls' different domain-specific abilities. Such beliefs, however, may have significant consequences on students' outcomes. Teachers acting upon gender stereotypes could—consciously or unconsciously—shape social interactions in class by, for example, creating a warm and challenging atmosphere for students from positively stereotyped groups and a cold and less challenging environment for students from negatively stereotyped groups (Aronson & Steele, 2005). Moreover, in one of the few studies on teachers' explicit gender stereotypes, Tiedemann (2002) found that stereotypes are related to the teachers' beliefs about effort and ability in mathematics.

In research into the effects of teachers' gender stereotypes, one should be aware of the particular age of the students participating in the investigation. There is some research showing that with increasing age, children become more and more aware of widely held stereotypes (Martinot, Bagès, & Désert, 2012; McKown & Weinstein, 2003) and are more likely to endorse traditional stereotypes themselves (Rowley, Kurtz-Costes, Mistry, & Feagans, 2007). Even more important, in relation to the present research is that students in late childhood or early adolescence become more and more aware of other persons' stereotypes. In a study by Kurtz-Costes, Rowley, Harris-Britt, and Woods (2008), for example, middle school children seemed to be more aware of adult stereotypes than were elementary school children. Thus, even though teachers may, of course, shape students' self-concepts at a young age, the focus of the present research on investigating the effects of teacher stereotypes in late childhood seemed appropriate.

The Present Investigation

Drawing on the idea that the gender-related beliefs and actions of significant others such as peers, parents, and teachers may affect the development of students' academic self-concept (e.g., Eccles & Wigfield, 2002; Wigfield & Eccles, 2000), we aimed to investigate

the relation of teachers' gender stereotypes to students' reading self-concept. We followed a longitudinal design with two waves of data collection. Our study went beyond previous research, as we investigated the consequences of teachers' explicit gender beliefs for the development of reading self-concept as a relatively stable personal characteristic. We analyzed the effect of teachers' stereotypes on reading self-concept over and above previous reading achievement. This is important, because it is obvious that prior academic achievement is influential in the formation of subsequent self-concept (Shavelson, Hubner, & Stanton, 1976); this is also true in the domain of reading (Retelsdorf, Köller, & Möller, 2014). Moreover, to account for the possible influence of ability grouping on students' self-concept (e.g., Marsh et al., 2008), we included the aggregated between-level achievement and reading self-concept as well as school track into our data analysis. In Germany, after elementary school, students are assigned to different types of school; these aim to prepare students either for a vocational apprenticeship (nonacademic track schools) or for university entrance (academic track schools).

Since gender beliefs about reading stereotypically favor girls (Plante et al., 2013; Schmenk, 2004), we expected that the negative gender stereotypes of boys' reading abilities would affect their reading self-concept. For girls, however, the expectations were less clear. On the one hand, there is evidence that even positive stereotypes can have negative effects, because high expectations may lead to so-called *choking under pressure*, which results in lower performance (Cheryan & Bodenhausen, 2000). On the other hand, girls' reading self-concepts are quite positive (Archambault et al., 2010), and the effects of stereotypes are generally expected to be rather small, so that a significant effect of teachers' gender stereotype on girls' reading self-concept was not expected.

Method

Sample and Procedure

Our sample stemmed from the larger longitudinal project LISA (in the German: *Lesen in der Sekundarstufe* [Reading in secondary school]), which mainly deals with the individual and contextual determinants of reading comprehension (e.g., Retelsdorf, Becker, Köller, & Möller, 2012; Retelsdorf, Köller, & Möller, 2011). This study drew on a sample of 1,508 secondary school students from 60 classes, drawn as representative of the federal state of Schleswig-Holstein, Germany. Data collection was performed by trained research students and took place as group tests carried out in class during regular lessons. The student questionnaire including the reading self-concept measure and the reading achievement tests was administered toward the beginning of Grade 5, a few weeks after the beginning of the school year (T1) and again after an interval of approximately 18 months, in the second half of Grade 6 (T2). Moreover, within 14 days of the data collection among the students at T1, all 60 German language teachers were asked to work on a teacher questionnaire including the items measuring their gender stereotypes; 54 teachers answered (66% female). Thereby, it is the established practice in secondary school that teachers usually change only every 2 years. In this study, only those students were included for whom teacher data also were available; this reduced the sample to 1,358 students (49% girls; girls' age at T1: $M = 10.96$, $SD = 0.61$; boys' age at T1: $M = 10.96$, $SD = 0.61$).

10.82, $SD = 0.51$; 36% at academic track schools). Applying t tests for reading achievement and reading self-concept and chi-square-tests for students' gender, we tested whether the excluded students differed in the study variables from the included students. None of these tests yielded significant results ($p \geq .135$).

Measures

Reading self-concept. We assessed reading self-concept with a subscale from the Habitual Reading Motivation Questionnaire (Möller & Bonerad, 2007) that comprises four items measuring students' evaluations of their own reading skills. Thus, the self-concept items refer to the comprehension of texts rather than to more basic reading skills (e.g., "Generally, understanding texts is easy for me"). Students rated their agreement with each item on a 4-point Likert-type scale anchored at 1 (*does not apply to me*) and 4 (*applies to me*). Cronbach's alpha measures were sufficient at both waves of data collection ($\alpha_{T1} = .74$, $\alpha_{T2} = .75$).

Teachers' gender stereotypes. Teachers at T1 were asked to answer three questions measuring their gender stereotypes about reading. They were asked if boys or girls read better, read more, and have more fun reading. Each answer was rated on a 5-point Likert-type scale, anchored at 1 (*boys much better/more*) and 5 (*girls much better/more*). The reliability of the scale was good ($\alpha = .87$).

Reading achievement. In this study, we used reading comprehension tests from the German section of the Progress in International Reading Literacy Study (Bos et al., 2005). The students' task was to read several texts and answer questions on their content. The questions mainly focused on students' skills in forming a broad and general understanding of the texts and in retrieving information from the texts. The test comprised 27 items—mainly multiple-choice items with four possible answers, but some open-format questions also were included. The item parameters were estimated by applying the partial credit model, because some items were scored polytomously. We estimated weighted-likelihood estimates (WLE) as subjects' ability scores using ConQuest (Wu, Adams, & Wilson, 1998). The WLE-reliability of the reading tests was sufficient (.82).

Statistical Analyses

We analyzed the association between teachers' gender stereotypes and students' self-concept by means of multiple group multilevel modeling, using Mplus Version 7.1 (Muthén & Muthén, 2013). Thereby, every teacher taught one class in our sample so that between-teacher and between-class effects are the same. Reading self-concept at T1, reading achievement, and teachers' gender stereotype were standardized ($M = 0$, $SD = 1$). Reading self-concept at T2 was standardized at the T1 mean and standard deviation of reading self-concept. To test our assumption that teachers' gender stereotypes affect boys' but not girls' self-concept, we specified a multiple group model with gender as a grouping variable. Since every teacher, however, teaches boys and girls, we had to deal with the situation that the grouping variable was within level. Thus, within each cluster, there could be varying random effects for boys and girls that could not be directly specified as multiple group multilevel models. Asparouhov and Muthén (2012) have suggested introducing latent variables that

represent this variation in between-level random effects. This approach also allows proper accounting for the covariance between the two group specific cluster effects. We tested a series of models predicting reading self-concept at T2 using this approach. In the first model, we included reading self-concept at T1 and teachers' gender stereotypes as predictors. In the second model, we additionally controlled for students' reading achievement at T1. Third, we additionally included aggregated scores of reading self-concept and reading achievement at T1 and school track as between-level covariates. The aggregated data were not standardized again at between-level.

We evaluated effect sizes to facilitate the interpretation of our results, following Tymms's (2004) proposal for calculating effect sizes in multilevel models. The effect size Δ can be interpreted similarly to Cohen's d (Cohen, 1988) and is calculated using the unstandardized regression coefficient in the multilevel model, the standard deviation of the predictor variable at between level, and the residual standard deviation at within level.

Due to missing data, we used multiple imputed data in all analyses as a state-of-the-art approach to address this problem (cf. Graham, 2009). On average, 11% of the data per variable were missing. We applied multiple imputation to create 20 complete data sets using Mplus 7.1 (see Graham, Olchowski, & Gilreath, 2007, for a discussion on the sufficient number of imputations). All subsequent analyses were then conducted 20 times, and the results were combined automatically in Mplus.

Results

Descriptive Statistics

As presented in Table 1, students' reading self-concept was above the theoretical mean of 2.5 at T1 and T2, indicating that students were quite confident in their reading skills. Moreover, boys had a higher reading self-concept than girls at T1, whereas girls had a higher reading self-concept than boys at T2. However, none of these differences yielded significance in a Wald chi-square test: $\chi^2(1) \leq 3.714$, $p \geq .054$. Girls also gained higher reading achievement scores at T1. Finally, the relatively high score of teachers' gender stereotypes indicated that, on average, the teachers believed that girls had higher reading abilities than boys.

Multilevel Analyses

We estimated the intraclass correlation (ICC) for reading self-concept at T2, testing the proportion of total variance that

Table 1

Means and Standard Deviations of the Study Variables

Variable	Overall		Girls		Boys	
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
Reading self-concept T1	3.03	0.70	2.99	0.71	3.08	0.68
Reading self-concept T2	3.08	0.63	3.10	0.61	3.06	0.65
Reading achievement T1	-0.05	1.12	0.01	1.02	-0.11	1.09
Teachers' gender stereotype	3.91	0.60				

Note. Weighted likelihood estimates have been estimated as subjects' ability scores for reading achievement. $N_{\text{teachers}} = 54$, $N_{\text{students}} = 1,358$. T1 = Time 1; T2 = Time 2.

could be attributed to between-class differences, resulting in an ICC of .114. Thus, with more than 10%, a substantial amount of the variance in reading self-concept goes back to differences between classes.

The results of our multiple group multilevel analyses are presented in Table 2. First, we tested a model (Model 1) in which reading self-concept at T1 was included as a within-level predictor and teachers' gender stereotype as a between-level predictor of reading self-concept at T2. For boys and girls, reading self-concept proved to be a significant predictor; thus indicating a certain stability of reading self-concept. Moreover, as expected, a significant negative effect of teachers' gender stereotypes on students' reading self-concept was recorded for boys but not for girls (effect sizes: $\Delta_{\text{boys}} = -.28$, $\Delta_{\text{girls}} = -.01$). The difference between boys and girls was tested by applying a Wald chi-square test, which indicated that the association between teachers' gender stereotype and reading self-concept was significantly stronger for boys than for girls, $\chi^2(1) = 11.05$, $p < .001$. In Model 2, we additionally included reading achievement at T1 as a within-level predictor; this also proved to be a significant predictor of reading self-concept at T2. The effect of reading self-concept at T1 was still significant but slightly smaller than in Model 1. Moreover, the negative effect of teachers' gender stereotype was again recorded for boys but not for girls (effect sizes: $\Delta_{\text{boys}} = -.25$, $\Delta_{\text{girls}} = -.03$). Again, this difference was significant, $\chi^2(1) = 6.10$, $p < .05$. Finally, we tested a third model (Model 3), in which we additionally included aggregate scores of reading achievement at T1 and reading self-concept at T1 and school track as between-level covariates. None of these additional variables yielded significance. Moreover, the effects of the within-level predictors and teachers' gender stereotype were similar to those in Model 2 (effect sizes: $\Delta_{\text{boys}} = -.23$, $\Delta_{\text{girls}} = -.04$). The Wald chi-square test comparing the effect of teachers' gender stereotype between boys and girls, again was significant, $\chi^2(1) = 3.94$, $p < .05$.

To illustrate the differential associations between teachers' gender stereotypes and students' reading self-concept, we plotted simple slopes for the results of Model 2 for boys and for girls (Figure 1). We chose this model because the additional predictors in Model 3 did not contribute to the prediction of reading self-concept at T2. Therefore, we understand Model 2 to be the most

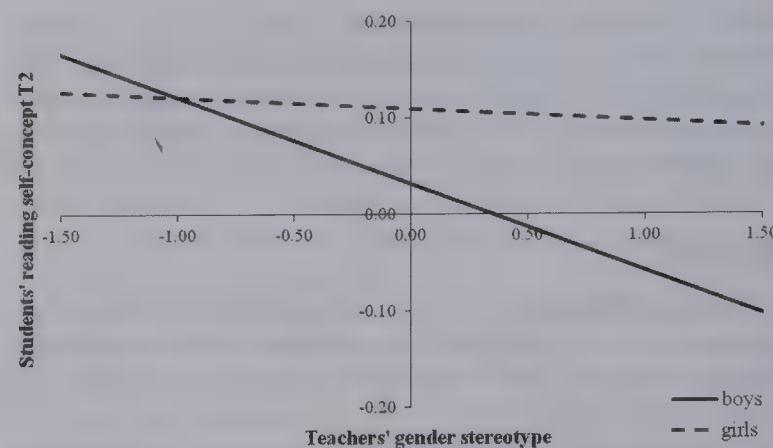


Figure 1. Relation between teachers' gender stereotype on boys' and girls' reading self-concept at T2 (from Model 2 in Table 1; all variables have been standardized). T2 = Time 2.

relevant model; it also meets the claim of parsimony. The simple slope analysis for Model 3, however, resulted in a similar pattern. Stronger gender stereotypes—such as, that teachers believe that girls outperform boys in reading—are associated with boys' lower reading self-concept, whereas girls' reading self-concept was unaffected by teachers' stereotype.

As an exploratory analysis, we also tested whether teachers' gender or the interaction Teachers' Gender \times Teachers' Gender Stereotype had different effects on boys' and girls' reading self-concept at T2. In line with the assumption of the so-called *same-sex teacher advantage* (for a detailed discussion, see Neugebauer, Helbig, & Landmann, 2011), one might have expected that boys' reading self-concept would benefit from a male teacher and girls' reading self-concept might benefit from a female teacher. Moreover, these benefits could be due to different gender stereotypes, depending on the teachers' gender. However, neither teachers' gender nor the interaction term was significant predictors of boys' and girls' reading self-concepts ($p \geq .154$). Moreover, the results of the model, including teachers' gender and their interaction, were by and large the same as the results of Model 3.

Table 2
Results of the Multiple Group Multilevel Analyses Predicting Reading Self-Concept at T2

Variable	Model 1				Model 2				Model 3			
	Girls		Boys		Girls		Boys		Girls		Boys	
	B	SE	B	SE	B	SE	B	SE	B	SE	B	SE
Within level												
Reading self-concept T1	.456	.030	.474	.035	.387	.032	.389	.033	.389	.033	.378	.035
Reading achievement T1					.243	.030	.273	.037	.257	.037	.254	.046
Between level												
Teachers' gender stereotype T1	-.003	.026	-.103	.035	-.012	.024	-.090	.033	-.013	.024	-.082	.034
Reading self-concept T1									.044	.156	.017	.149
Reading achievement T1									-.053	.095	-.006	.106
School track									-.019	.088	.113	.096

Note. All variables but the dummies have been standardized (reading self-concept T2 was standardized at the mean at standard deviation of reading self-concept T1); school track was dummy-coded (0 = nonacademic track, 1 = academic track). $N_{\text{teachers}} = 54$, $N_{\text{students}} = 1,358$. Bold = parameters are significant ($p < .05$). T1 = Time 1; T2 = Time 2.

Discussion

The aim of this research was to investigate whether teachers' stereotypes affected students' self-concepts in reading, a stereotypically female domain. We expected teachers' gender stereotypes about students' reading abilities—namely, that girls perform better in reading tasks—to negatively affect boys' but not girls' reading self-concepts. Therefore, we drew on longitudinal data comprising two waves of data collection to predict students' reading self-concept at the end of Grade 6 with the previously reported (at the beginning of Grade 5) teacher stereotypes, controlling for previous reading self-concept. Our hypothesis was corroborated: boys' reading self-concept in Grade 6 was lower for students whose teachers reported high scores for gender stereotypes. No effect was recorded for girls. Moreover, the effect was also robust when students' previous achievement on individual and class levels, and their school track were included. Thus, our results have shown that teachers' gender stereotypes negatively affect boys' reading self-concept over and above their actual performance. Additionally, our results indicate that, on average, teachers' reading stereotypes favor girls. Consequently it is possible that even less stereotyped teachers favor girls over boys in the reading domain, indicating that the total effect of gender stereotypes might be greater than we can show in our analyses. However, this interpretation is rather speculative and needs further research.

Before discussing the implications of our findings in more detail, we first discuss the absence of gender differences in the mean level of reading self-concept. This finding is in line with other research that does not support the assumption of gender differences in language-related self-concepts (Anderman et al., 2001; Evans et al., 2011; Skaalvik & Skaalvik, 2004). One possible explanation deals with the particular age of our students. Conjecturally, at the onset of puberty, the intensification of gender differences is only beginning. Although there is no evidence for such gender intensification in longitudinal studies (Jacobs et al., 2002), our results do suggest the tendency for an opposing trend in girls' and boys' reading self-concept, favoring girls. Another explanation, provided by Skaalvik and Skaalvik (2004), deals with the idea that gender differences in self-concepts are based on perceptions of individual strength and weaknesses across different domains—similar to what is proposed in the dimensional comparison theory (Möller & Marsh, 2013). Thus, gender differences would not become obvious in group comparisons within a single domain but only in comparisons of self-concepts in different domains. Our data however do not allow for analyses along these lines. Regardless of this question of group differences, our results nevertheless provide some evidence that variability in reading self-concept development may be explained in part by teachers' gender stereotypes. In the remainder of this article, we discuss the implications of these findings.

Theoretical and Practical Implications

Our findings help our understanding of the development of reading self-concept in secondary school and contribute to our knowledge of possible reasons for gender differences in self-concept. However, even though our study comprised longitudinal data, and we were able to control for important predictors of reading self-concept, we cannot draw causal conclusions, since we cannot rule out the effects of unobserved variables. Our study,

however, complements experimental data on the consequences of specific stereotype content (e.g., Becker & Asbrock, 2012; Cuddy et al., 2007) by providing high external validity, due to the naturalistic setting in actual school life. The results support the assumption of expectancy-value theory, that gender beliefs of significant others play an important role in shaping students' ability self-concepts. We found evidence that, in addition to parents' (Jacobs & Eccles, 1992; Tiedemann, 2000) and students' own (Plante et al., 2013) stereotypes, teachers' gender stereotypes also play an important role in shaping students' self-concepts—over and above students' actual achievement. Consequently, these stereotypes might explain to some extent why gender differences in language-related self-concept increase over time (Archambault et al., 2010; Jacobs et al., 2002). However, these effects were rather small in terms of Cohen's (1988) classification of effect sizes. At least in children 10 years and older, however, reading self-concept seems to be quite stable (Retelsdorf et al., 2014), so that large effects cannot be expected, and thus, even small effects may still be of practical relevance. This might be even more relevant when taking into account that teachers' stereotypes are a rather distal determinant of students' self-concept compared with their achievement or other student-level variables. It might be interesting, though, to test the relations of teachers' gender stereotypes with younger students' self-concept development. Jacobs et al. (2002) reported much greater decreases of language self-concept from Grade 1 to Grade 5 than from Grade 6 to Grade 12—particularly for boys. Thus, there might be sensitive developmental stages in which environmental influences have particularly pronounced effects on children's self-concept—however, younger students may not yet be aware of teachers' stereotypes (e.g., Muzzatti & Agnoli, 2007). Moreover, there may be greater cumulative effects over a longer period of time.

Another open question deals with the mediating processes. We were not able to investigate such processes between teachers' gender stereotypes and students' reading self-concept. Thus, we do not know whether teachers who think that boys are less able to read than girls actually treat boys and girls differently. However, it seems plausible that teachers' beliefs would influence their own behavior in classroom, as indicated by experimental studies on the effects of specific stereotypes on outgroup-directed behavior (e.g., Becker & Asbrock, 2012; Cuddy et al., 2007; for an overview, see Cuddy, Glick, & Beninger, 2011). As a consequence, boys' reading self-concept might suffer from teachers' behavior even when their reading abilities are similar to girls' abilities. As Rubie-Davies, Hattie, and Hamilton (2006) discussed, there is some evidence that teachers who hold stereotypes regarding particular groups alter their practices and limit opportunities to learn for negatively stereotyped students. This is in line with research on the effects of incompetence stereotypes: Groups perceived as incompetent are ignored or excluded more than other groups (Cuddy et al., 2011). Moreover, teachers believing in a certain stereotype may tend to make remarks or to behave in ways that make these stereotypes more salient in class, thus indicating stereotype threat (Aronson & Steele, 2005). Apart from teachers' classroom behavior, increasing awareness of widely held stereotypes (McKown & Weinstein, 2003) and developing knowledge of adults' stereotypes (Muzzatti & Agnoli, 2007) may shape the students' own gender beliefs. As a further consequence, they may react by adapting their

own self-concept to these gender beliefs (cf. Kurtz-Costes et al., 2008).

Another question deals with the problem of the accuracy of teachers' gender beliefs: the so-called "kernel of truth" of their gender stereotype. It could be argued that, taking into account recent results from large-scale assessments (cf. Mullis, Martin, Foy, & Drucker, 2012; Organization for Economic Co-Operation and Development, 2010), teachers' beliefs of girls outperforming boys in reading are to some extent true. This problem is somehow connected to the argument that teacher expectations have an impact on students' achievement simply because their expectations are accurate (Jussim & Harber, 2005). According to our results, this would mean that there is a negative effect of teachers' gender stereotype on boys' reading self-concept just because boys do in fact have lower reading abilities. These lower abilities should then lead to boys' decreasing reading self-concept. Similarly, since boys are likely to show declining motivation related to language-related tasks in Grade 5 and Grade 6 (cf. Jacobs et al., 2002), our results may reflect teachers' accurate appraisal of boys' declining language-related motivation. In this study, however, we controlled for individual achievement as well as for mean class achievement, so that teachers' gender stereotypes have been shown to exert an additional effect on students' reading self-concept that goes beyond the effect of actual achievement. Moreover, the teacher questions on gender stereotypes were generally worded, not related to the particular classes they were teaching. Considering that the teachers completed the questionnaire only a few weeks after they first met their students, it seems plausible to assume that the teachers' beliefs about gender differences were not affected by the individual students' motivational declines.

An important question that arises from our findings is what teachers can do to counteract the reported relation between their own stereotypes and boys' reading self-concept. Generally, it is a good idea to counteract prior gender stereotypes and make the expectation clear in class that boys and girls perform equally well (Hartley & Sutton, 2013). Moreover, during their teacher education, teachers should be apprised of the fact that their beliefs do have consequences and that, consciously or not, they may be prone to certain biases in their treatment of boys and girls. Although cultural stereotypes are widely shared and guide behavioral reactions, people can choose to overcome this automatic effect (Fiske, 2004). Most research investigating similar discriminatory behavior in class has dealt with girls in mathematics and science, and thus the question here is whether language teachers behave similarly. We cannot answer this question yet, due to the lack of research in language teaching, but it would appear that certain rules for teachers' classroom behavior, as summarized by Woolfolk (2010), should be introduced in the near future.

However, it should be noted that there is strong evidence that in general, teachers interact more frequently with boys than with girls (Jones & Dindia, 2004). This difference has mainly been found in relation to negative interactions such as criticism, while no difference in positive interactions such as praise or acceptance has been found. Unfortunately, Jones and Dindia did not test the effects of domain or school subject, so their meta-analysis does not provide information on differences in mathematics and language teaching. In a study by Worrall and Tsarna (1987), the self-reported classroom interactions of science and language teachers were compared. These authors found that girls are relatively favored in

language subjects, compared with boys, whereas no differences in science have been reported. Regarding the question as to what teachers can do, Woolfolk (2010) suggested a kind of checklist on how to avoid discriminatory behavior in the classroom. First of all, she encouraged teachers to be aware of bias in their own behavior. Do the teachers group boys and girls for certain tasks? Do they prefer boys or girls when asking questions regarding particular topics—for example, boys for technical and girls for social issues? Second, she asked teachers to check their teaching material for gender inequalities, such as presenting traditional role models. Third, teachers should have a critical look at general inequalities at the school—for example, if there is biased advice regarding course selection. Fourth, teachers should use gender-neutral language whenever possible. Fifth, teachers should introduce role models that do not represent traditional gender roles.

Conclusion

Our study complements previous research by investigating the effects of teachers' stereotypes on students' reading self-concept, drawing on a relatively large sample tested in a naturalistic setting. Our results suggest that not only do gender stereotypes have short-term effects like those investigated in the framework of stereotype threat theory (cf. Aronson & Steele, 2005), but they can also explain the long-term development of reading self-concept as a relatively stable personal characteristic. In our study, boys were the disadvantaged group. Therefore, we would like to follow Hartley and Sutton (2013) in noting that these results have to be considered in the light of general male advantage in society, such as the gender pay gap that still persists (e.g., Council of the European Union, 2010; Drago & Williams, 2010). However, it should not be the aim to pit males' advantages in one area against their disadvantages in another area. We should encourage and enable our teachers to counteract prior gender stereotypes and to become aware of their own potentially discriminatory behaviors. One important condition for an equitable educational system is that teachers should become aware of and resistant to stereotypes.

References

- Anderman, E. M., Eccles, J. S., Yoon, K. S., Roeser, R. W., Wigfield, A., & Blumenfeld, P. C. (2001). Learning to value mathematics and reading: Relations to mastery and performance-oriented instructional practices. *Contemporary Educational Psychology*, 26, 76–95. doi:10.1006/ceps.1999.1043
- Archambault, I., Eccles, J. S., & Vida, M. N. (2010). Ability self-concepts and subjective value in literacy. Joint trajectories from Grades 1 through 12. *Journal of Educational Psychology*, 102, 804–816. doi:10.1037/a0021075
- Aronson, J. M., & Steele, C. M. (2005). Stereotypes and the fragility of academic competence, motivation, and self-concept. In A. J. Elliot & C. S. Dweck (Eds.), *Handbook of competence and motivation* (pp. 436–456). New York, NY: Guilford Press.
- Asbrock, F., Nieuwoudt, C., Duckitt, J., & Sibley, C. G. (2011). Societal stereotypes and the legitimization of intergroup behavior in Germany and New Zealand. *Analysis of Social Issues and Public Policy*, 11, 154–179. doi:10.1111/j.1530-2415.2011.01242.x
- Asparouhov, T., & Muthén, B. O. (2012). *Multiple group multilevel analysis* (Mplus Web Notes, No. 16). Retrieved from www.statmodel.com/examples/webnotes/webnote16.pdf
- Becker, J. C., & Asbrock, F. (2012). What triggers helping versus harming of ambivalent groups? Effects of the relative salience of warmth versus

competence. *Journal of Experimental Social Psychology*, 48, 19–27. doi:10.1016/j.jesp.2011.06.015

Bos, W., Lankes, E.-M., Prenzel, M., Schwippert, K., Valtin, R., Voss, A., & Walther, G. (2005). *IGLU: Skalenhandbuch zur Dokumentation der Erhebungsinstrumente*. [IGLU: Scale manual for documentation of the survey instruments in the PIRLS (Progress in International Reading Literacy Study)]. Münster, Germany: Waxmann.

Cheryan, S., & Bodenhausen, G. V. (2000). When positive stereotypes threaten intellectual performance: The psychological hazards of "model minority" status. *Psychological Science*, 11, 399–402. doi:10.1111/1467-9280.00277

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Erlbaum.

Council of the European Union. (2010). *The gender pay gap in the member states of the European Union: Quantitative and qualitative indicators. Summary of the Belgian presidency report 2010*. Brussels, Belgium: Author. Retrieved from <http://register.consilium.europa.eu/pdf/en/10/st16/st16881-ad01.en10.pdf>

Cuddy, A. J. C., Fiske, S. T., & Glick, P. (2007). The BIAS map: Behaviors form Intergroup Affect and Stereotypes. *Journal of Personality and Social Psychology*, 92, 631–648. doi:10.1037/0022-3514.92.4.631

Cuddy, A. J. C., Glick, P., & Beninger, A. (2011). The dynamics of warmth and competence judgments, and their outcomes in organizations. *Research in Organizational Behavior*, 31, 73–98. doi:10.1016/j.riob.2011.10.004

de Boer, H., Bosker, R. J., & van der Werf, M. P. C. (2010). Sustainability of teacher expectation bias effects on long-term student performance. *Journal of Educational Psychology*, 102, 168–179. doi:10.1037/a0017289

DeZolt, D. M., & Hull, S. H. (2001). Classroom and school climate. In J. Worell (Ed.), *Encyclopedia of women and gender*. Sex similarities and differences and the impact of society on gender (pp. 257–264). San Diego, CA: Academic Press.

Drago, R., & Williams, C. (2010). *The gender wage gap 2009*. Washington, DC: Institute for Women's Policy Research. Retrieved from www.iwpr.org/publications/pubs/the-gender-wage-gap-2009/at_download/file

Durik, A. M., Vida, M., & Eccles, J. S. (2006). Task values and ability beliefs as predictors of high school literacy choices: A developmental analysis. *Journal of Educational Psychology*, 98, 382–393. doi:10.1037/0022-0663.98.2.382

Eccles, J. S., Adler, T. F., Futterman, R., Goff, S. B., Kaczala, C. M., Meece, J. L., & Midgley, C. (1983). Expectations, values, and academic behaviors. In J. T. Spence (Ed.), *Achievement and achievement motives: Psychological and sociological approaches* (pp. 75–146). San Francisco, CA: Freeman.

Eccles, J. S., & Wigfield, A. (2002). Motivational beliefs, values, and goals. *Annual Review of Psychology*, 53, 109–132. doi:10.1146/annurev.psych.53.100901.135153

Eccles, J. S., Wigfield, A., Harold, R. D., & Blumenfeld, P. C. (1993). Age and gender differences in children's self- and task perceptions during elementary school. *Child Development*, 64, 830–847. doi:10.2307/1131221

Evans, A. B., Coping, K. E., Rowley, S. J., & Kurtz-Costes, B. (2011). Academic self-concept in Black adolescents: Do race and gender stereotypes matter? *Self and Identity*, 10, 263–277. doi:10.1080/15298868.2010.485358

Fiedler, K., & Bless, H. (2001). Social cognition. In M. Hewstone & W. Stroebe (Eds.), *Introduction to social psychology: A European perspective* (pp. 115–149). Malden, MA: Blackwell.

Fiske, S. T. (2004). What's in a category? Responsibility, intent, and the avoidability of bias against outgroups. In A. G. Miller (Ed.), *The social psychology of good and evil* (pp. 127–140). New York, NY: Guilford Press.

Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual Review of Psychology*, 60, 549–576. doi:10.1146/annurev.psych.58.110405.085530

Graham, J. W., Olchowski, A. E., & Gilreath, T. D. (2007). How many imputations are really needed? Some practical clarifications of multiple imputation theory. *Prevention Science*, 8, 206–213. doi:10.1007/s11121-007-0070-9

Hartley, B. L., & Sutton, R. M. (2013). A stereotype threat account of boys' academic underachievement. *Child Development*, 84, 1716–1733. doi:10.1111/cdev.12079

Hyde, J. S., & Durik, A. M. (2005). Gender, competence, and motivation. In A. J. Elliot & C. S. Dweck (Eds.), *Handbook of competence and motivation* (pp. 375–391). New York, NY: Guilford Press.

Ireson, J., & Hallam, S. (2009). Academic self-concepts in adolescence: Relations with achievement and ability grouping in schools. *Learning and Instruction*, 19, 201–213. doi:10.1016/j.learninstruc.2008.04.001

Jacobs, J. E., & Eccles, J. S. (1992). The impact of mothers' gender-role stereotypic beliefs on mothers' and children's ability perceptions. *Journal of Personality and Social Psychology*, 63, 932–944. doi:10.1037/0022-3514.63.6.932

Jacobs, J. E., Lanza, S., Osgood, D. W., Eccles, J. S., & Wigfield, A. (2002). Changes in children's self-competence and values: Gender and domain differences across Grades One through Twelve. *Child Development*, 73, 509–527. doi:10.1111/1467-8624.00421

Jones, S. M., & Dindia, K. (2004). A meta-analytic perspective on sex equity in the classroom. *Review of Educational Research*, 74, 443–471. doi:10.3102/00346543074004443

Jussim, L., Eccles, J. S., & Madon, S. (1996). Social perception, social stereotypes, and teacher expectations: Accuracy and the quest for the powerful self-fulfilling prophecy. In M. P. Zanna (Ed.), *Advances in experimental social psychology* (pp. 281–388). San Diego, CA: Academic Press.

Jussim, L., & Harber, K. D. (2005). Teacher expectations and self-fulfilling prophecies: Knowns and unknowns, resolved and unresolved controversies. *Personality and Social Psychology Review*, 9, 131–155. doi:10.1207/s15327957pspr0902_3

Kurtz-Costes, B., Rowley, S. J., Harris-Britt, A., & Woods, T. A. (2008). Gender stereotypes about mathematics and science and self-perceptions of ability in late childhood and early adolescence. *Merrill-Palmer Quarterly*, 54, 386–409. doi:10.1353/mpq.0.0001

Marsh, H. W., Seaton, M., Trautwein, U., Lüdtke, O., Hau, K.-T., O'Mara, A. J., & Craven, R. G. (2008). The big-fish-little-pond-effect stands up to critical scrutiny: Implications for theory, methodology, and future research. *Educational Psychology Review*, 20, 319–350. doi:10.1007/s10648-008-9075-6

Marsh, H. W., Trautwein, U., Lüdtke, O., Köller, O., & Baumert, J. (2006). Integration of multidimensional self-concept and core personality constructs: Construct validation and relations to well-being and achievement. *Journal of Personality*, 74, 403–456. doi:10.1111/j.1467-6494.2005.00380.x

Martinot, D., Bagès, C., & Désert, M. (2012). French children's awareness of gender stereotypes about mathematics and reading: When girls improve their reputation in math. *Sex Roles*, 66, 210–219. doi:10.1007/s11199-011-0032-3

McKown, C., & Weinstein, R. S. (2003). The development and consequences of stereotype consciousness in middle childhood. *Child Development*, 74, 498–515. doi:10.1111/1467-8624.7402012

Meece, J. L., Bower Glienke, B., & Burg, S. (2006). Gender and motivation. *Journal of School Psychology*, 44, 351–373. doi:10.1016/j.jsp.2006.04.004

Möller, J., & Bonerad, E.-M. (2007). Fragebogen zur habituellen Lesemotivation. [Habitual Reading Motivation Questionnaire]. *Psychologie in Erziehung und Unterricht*, 54, 259–267.

Möller, J., & Marsh, H. W. (2013). Dimensional comparison theory. *Psychological Review, 120*, 544–560. doi:10.1037/a0032459

Möller, J., Retelsdorf, J., Köller, O., & Marsh, H. W. (2011). The reciprocal internal/external frame of reference model: An integration of models of relations between academic achievement and self-concept. *American Educational Research Journal, 48*, 1315–1346. doi:10.3102/0002831211419649

Mullis, I. V. S., Martin, M. O., Foy, P., & Drucker, K. T. (2012). *PIRLS 2011 international results in reading*. Chestnut Hill, MA: Boston College.

Muthén, L. K., & Muthén, B. O. (2013). Mplus Version 7.1 [Computer software]. Los Angeles, CA: Muthén & Muthén.

Muzzatti, B., & Agnoli, F. (2007). Gender and mathematics: Attitudes and stereotype threat susceptibility in Italian children. *Developmental Psychology, 43*, 747–759. doi:10.1037/0012-1649.43.3.747

Neugebauer, M., Helbig, M., & Landmann, A. (2011). Unmasking the myth of the same-sex teacher advantage. *European Sociological Review, 27*, 669–689. doi:10.1093/esr/jcq038

Nguyen, H.-H. D., & Ryan, A. M. (2008). Does stereotype threat affect test performance of minorities and women? A meta-analysis of experimental evidence. *Journal of Applied Psychology, 93*, 1314–1334. doi:10.1037/a0012702

Organization for Economic Co-Operation and Development. (2010). *PISA 2009 results: What students know and can do. Student performance in reading, mathematics and science* (Vol. 1). Paris, FranceAuthorhttp://dx.doi.org/10.1787/9789264091450-en

Plante, I., de la Sablonnière, R., Aronson, J. M., & Théorêt, M. (2013). Gender stereotype endorsement and achievement-related outcomes: The role of competence beliefs and task values. *Contemporary Educational Psychology, 38*, 225–235. doi:10.1016/j.cedpsych.2013.03.004

Retelsdorf, J., Becker, M., Köller, O., & Möller, J. (2012). Reading development in a tracked school system: A longitudinal study over 3 years using propensity score matching. *British Journal of Educational Psychology, 82*, 647–671. doi:10.1111/j.2044-8279.2011.02051.x

Retelsdorf, J., Köller, O., & Möller, J. (2011). On the effects of motivation on reading performance growth in secondary school. *Learning and Instruction, 21*, 550–559. doi:10.1016/j.learninstruc.2010.11.001

Retelsdorf, J., Köller, O., & Möller, J. (2014). Reading achievement and reading self-concept: Testing the reciprocal effects model. *Learning and Instruction, 29*, 21–30. doi:10.1016/j.learninstruc.2013.07.004

Rouland, K. K., Rowley, S. J., & Kurtz-Costes, B. (2013). Self-views of African-American youth are related to the gender stereotypes and ability attributions of their parents. *Self and Identity, 12*, 382–399. doi:10.1080/15298868.2012.682360

Rowley, S. J., Kurtz-Costes, B., Mistry, R., & Feagans, L. (2007). Social status as a predictor of race and gender stereotypes in late childhood and early adolescence. *Social Development, 16*, 150–168. doi:10.1111/j.1467-9507.2007.00376.x

Rubie-Davies, C., Hattie, J. A. C., & Hamilton, R. (2006). Expecting the best for students: Teacher expectations and academic outcomes. *British Journal of Educational Psychology, 76*, 429–444. doi:10.1348/000709905X53589

Schmenk, B. (2004). Language learning: A feminine domain? The role of stereotyping in constructing gendered learner identities. *TESOL Quarterly, 38*, 514–524. doi:10.2307/3588352

Schneider, D. J. (2004). *The psychology of stereotyping*. New York, NY: Guilford Press.

Shavelson, R. J., Hubner, J. J., & Stanton, G. C. (1976). Self-concept: Validation of construct interpretations. *Review of Educational Research, 46*, 407–441. doi:10.3102/00346543046003407

Skaalvik, S., & Skaalvik, E. M. (2004). Gender differences in math and verbal self-concept, performance expectations, and motivation. *Sex Roles, 50*, 241–252. doi:10.1023/B:SERS.0000015555.40976.e6

Steele, C. M. (1997). A threat in the air: How stereotypes shape intellectual identity and performance. *American Psychologist, 52*, 613–629. doi:10.1037/0003-066X.52.6.613

Swinson, J., & Harrop, A. (2009). Teacher talk directed to boys and girls and its relationship to their behaviour. *Educational Studies, 35*, 515–524. doi:10.1080/03055690902883913

Tajfel, H. (1981). Social stereotypes and social groups. In J. C. Turner & H. Giles (Eds.), *Intergroup behavior* (pp. 144–167). Oxford, England: Blackwell.

Tajfel, H., & Turner, J. C. (1986). The social identity theory of intergroup behavior. In S. Worchel & W. G. Austin (Eds.), *Psychology of intergroup relations* (pp. 7–24). Chicago, IL: Nelson-Hall.

Tiedemann, J. (2000). Parents' gender stereotypes and teachers' beliefs as predictors of children's concept of their mathematical ability in elementary school. *Journal of Educational Psychology, 92*, 144–151. doi:10.1037/0022-0663.92.1.144

Tiedemann, J. (2002). Teachers' gender stereotypes as determinants of teacher perceptions in elementary school mathematics. *Educational Studies in Mathematics, 50*, 49–62. doi:10.1023/A:1020518104346

Tymms, P. (2004). Effect sizes in multilevel models. In I. Schagen & K. Elliot (Eds.), *But what does it mean? The use of effect sizes in educational research* (pp. 55–66). London, England: National Foundation for Educational Research.

Watt, H. M. G., & Eccles, J. S. (Eds.). (2008). *Gender and occupational outcomes: Longitudinal assessments of individual, social, and cultural influences*. Washington, D.C: American Psychological Association. doi:10.1037/11706-000

Wigfield, A., & Eccles, J. S. (2000). Expectancy-value theory of achievement motivation. *Contemporary Educational Psychology, 25*, 68–81. doi:10.1006/ceps.1999.1015

Wigfield, A., Eccles, J. S., Yoon, K. S., Harold, R. D., Arbreton, A. J. A., Freedman-Doan, C., & Blumenfeld, P. C. (1997). Change in children's competence beliefs and subjective task values across the elementary school years: A 3-year study. *Journal of Educational Psychology, 89*, 451–469. doi:10.1037/0022-0663.89.3.451

Woolfolk, A. E. (2010). *Educational psychology*. Columbus, OH: Pearson/Allyn & Bacon.

Worrall, N., & Tsarna, H. (1987). Teachers' reported practices towards girls and boys in science and languages. *British Journal of Educational Psychology, 57*, 300–312. doi:10.1111/j.2044-8279.1987.tb00859.x

Wu, M. L., Adams, R. J., & Wilson, M. (1998). ConQuest: Generalized item response modeling software [Computer software]. Melbourne, Australia: Australian Council for Educational Research.

Received October 27, 2013
Revision received April 29, 2014

Accepted May 3, 2014 ■

Gender Differences in the Effects of a Utility-Value Intervention to Help Parents Motivate Adolescents in Mathematics and Science

Christopher S. Rozek, Janet S. Hyde,
and Ryan C. Svoboda
University of Wisconsin—Madison

Chris S. Hulleman
University of Virginia

Judith M. Harackiewicz
University of Wisconsin—Madison

A foundation in science, technology, engineering, and mathematics (STEM) education is critical for students' college and career advancement, but many U.S. students fail to take advanced mathematics and science classes in high school. Research has neglected the potential role of parents in enhancing students' motivation for pursuing STEM courses. Previous research has shown that parents' values and expectancies may be associated with student motivation, but little research has assessed the influence of parents on adolescents through randomized experiments. Harackiewicz, Rozek, Hulleman, and Hyde (2012) documented an increase in adolescents' STEM course-taking for students whose parents were assigned to a utility-value intervention in comparison to a control group. In this study, we examined whether that intervention was equally effective for boys and girls and examined factors that moderate and mediate the effect of the intervention on adolescent outcomes. The intervention was most effective in increasing STEM course-taking for high-achieving daughters and low-achieving sons, whereas the intervention did not help low-achieving daughters (prior achievement measured in terms of grade point average in 9th-grade STEM courses). Mediation analyses showed that changes in STEM utility value for mothers and adolescents mediated the effect of the intervention on 12th-grade STEM course-taking. These results are consistent with a model in which parents' utility value plays a causal role in affecting adolescents' achievement behavior in the STEM domain. The findings also indicate that utility-value interventions with parents can be effective for low-achieving boys and for high-achieving girls but suggest modifications in their use with low-achieving girls.

Keywords: academic motivation, educational intervention, STEM motivation, gender differences

In the United States, national education policies have focused on improving the performance of U.S. students relative to their international peers, particularly in areas related to science, technology, engineering, and mathematics (STEM; National Science Foundation [NSF], 2012). Of particular concern are students' decisions not to take advanced science and mathematics courses in high school. For example, only 35% of high school graduates have taken precalculus and only 39% have taken physics (NSF, 2012). Moreover, although gender gaps have closed for course-taking in some STEM areas, they persist in others. For example, although

girls and boys take calculus at the same rate, boys are more likely to take physics than girls are (42% vs. 36%) and are more likely to take engineering in high school (6% vs. 1%; NSF, 2012). Recently, a number of interventions have been implemented to increase STEM motivation and to close gender gaps (e.g., Harackiewicz et al., 2014; Hulleman & Harackiewicz, 2009; Miyake et al., 2010; Walton & Cohen, 2011). Here, we report on the moderators and mediators of an intervention shown to help parents motivate their adolescents to take mathematics and science courses in high school (Harackiewicz, Rozek, Hulleman, & Hyde, 2012). We probed

This article was published Online First June 2, 2014.

Christopher S. Rozek, Janet S. Hyde, and Ryan C. Svoboda, Department of Psychology, University of Wisconsin—Madison; Chris S. Hulleman, Center for Advanced Study of Teaching and Learning, University of Virginia; Judith M. Harackiewicz, Department of Psychology, University of Wisconsin—Madison.

This research was supported by the National Science Foundation (Grant DRL 0814750) and by the Institute for Education Sciences, U.S. Department of Education, through Award No. R305B090009 and Grant 144-NL14 to the University of Wisconsin—Madison. The opinions expressed are those of the authors and do not represent views of the National Science Foundation or the U.S. Department of Education. We thank Carlie Allison,

Corinne Boldt, Andrew Carpenter, Claire Johnson, Kerstin Krautbauer, Dan Lamanna, Anita Lee, Maria Mens, Michael Noh, Jenni Petersen, Margaret Wolfram, and Justin Wu for their help conducting this research, and we thank the members of our advisory board (Jacque Eccles, Adam Gamoran, Jo Handelsman, Jenefer Husman, Dominic Johann-Berkel, Ann Renninger, and Judith Smetana) for their guidance. We are most grateful to the families of the Wisconsin Study of Families and Work project for their participation over the years.

Correspondence concerning this article should be addressed to Janet S. Hyde, Department of Psychology, University of Wisconsin—Madison, 1202 West Johnson Street, Madison, WI 53706-1611. E-mail: jshyde@wisc.edu

whether the intervention was equally effective for boys and girls depending on their prior performance in mathematics and science courses and what factors mediated the effect of the intervention on students' STEM course-taking.

Theoretical Framework

Numerous theoretical models have been proposed to help explain student motivation and persistence in academics. One comprehensive model is Eccles's expectancy-value theory (Eccles-Parsons et al., 1983), which frames the research reported here. The expectancy-value model holds that expectations for success (expectancy) and perceived task value are direct predictors of achievement and achievement choices (e.g., Eccles-Parsons et al., 1983; Simpkins, Davis-Kean, & Eccles, 2006; Updegraff, Eccles, Barber, & O'Brien, 1996). In Eccles's model, expectancy for success is defined as how well an individual thinks he or she will do on an ensuing task (Eccles-Parsons et al., 1983). Task value consists of attainment value (how a task is related to one's identity), intrinsic value (enjoyment of the task), utility value (perceived usefulness of a task), and cost (costs to the individual of task engagement, such as what one concedes by choosing one task over others).

The expectancy-value model proposes that adolescents' perceived task values and expectations for success are the most proximal predictors of STEM-related achievement choices. Previous research supports this hypothesis, with students being more likely to choose to take mathematics and science courses when they have either high expectations for success or value for those courses or both (e.g., Eccles, Barber, Updegraff, & O'Brien, 1998; Simpkins et al., 2006; Updegraff et al., 1996; Watt, 2005; Watt, Eccles, & Durik, 2006). In addition, both expectancies and values predict classroom performance (e.g., Hulleman, Durik, Schweigert, & Harackiewicz, 2008; Watt, 2005).

Parents' Influence on Values and Expectancies

The expectancy-value model proposes that, more distally, key socializers, such as parents, play an important role in shaping adolescents' values. Previous research has found that parents' values and expectancies for success for their child are linked to adolescents' values in a variety of domains, including mathematics and science (Jodl, Michael, Malanchuk, Eccles, & Sameroff, 2001; Simpkins, Fredericks, & Eccles, 2012). Much of this research has concentrated on adolescents and their achievement motivation in STEM courses throughout middle school and high school (Riegle-Crumb & King, 2010; Watt et al., 2012). Parents' values for mathematics and science are associated with adolescents' values in mathematics and science, which, subsequently, are associated with adolescents' educational choices and outcomes (Jodl et al., 2001; Simpkins et al., 2012).

Parents' expectancies for their adolescents have also been associated with their adolescents' expectancies for success in mathematics and science and educational outcomes, and these associations are even stronger than the associations between parents' values and adolescents' outcomes (Bleeker & Jacobs, 2004; Frome & Eccles, 1998; Jacobs & Eccles, 1992; Yee & Eccles, 1988). For instance, if parents have high expectancies for their adolescents in STEM, they are more likely to have adolescents with high expectancies and better educational outcomes in STEM courses. If parents have low expectancies, they are more likely to have adolescents with low expectancies and worse educational outcomes in STEM (Jacobs & Eccles, 1992). However, studies involving the associations between parents' and adolescents' expectations and values are typically correlational in nature and thus are unable to test for a causal effect of parents' values and expectations on adolescents' values and expectations.

Whereas multiple studies have focused on the role of parental support—such as involvement and support for autonomy—in relation to children's school outcomes (Grodnick & Ryan, 1989; Grodnick, Ryan, & Deci, 1991; Ratelle, Larose, Guay, & Senécal, 2005; Spera, 2005), here we focus on parents' values for their child's education. Such values may be a key resource that educators can leverage to enhance student outcomes, such as STEM course-taking (Harackiewicz et al., 2012). From a process perspective, it is important to understand how parents' values are transmitted to children. Some researchers have examined the specific parental behaviors that contribute to value transmission from parents to adolescents, such as encouragement, provision of educational and other materials, and coactivity (e.g., Simpkins et al., 2012). However, parental behaviors are not the only means of value transmission. Because students' perceptions are featured heavily in the expectancy-value model, it is important to examine whether adolescents are even aware of their parents' values. If adolescents are unaware of their parents' utility-value beliefs, parents' values may have smaller effects on their adolescents' attitudes and behaviors. Such perceptions could serve as an important indicator that parental values are being communicated (Paulson & Sputa, 1996; Spera, 2006; Wood, Kurtz-Costes, & Copping, 2011).

Gender Differences in Expectancies and Values

Two aspects of Eccles's model have been hypothesized to show gender differences that, in turn, may explain differences in STEM achievement: gender differences in expectancies and gender differences in values (e.g., Eccles, Wigfield, Harold, & Blumenfeld, 1993; Updegraff et al., 1996). Compared with boys, girls have lower expectancies for success in STEM domains (Yee & Eccles, 1988). This difference predicts increased enrollment in these courses for boys (Watt et al., 2012). Gender differences in expectancies for success can be influenced by socializers, especially parents. Research indicates that parents can have exaggerated expectancies for success in mathematics and science for their sons and diminished expectancies for success for their daughters (Eccles et al., 1993; Yee & Eccles, 1988).

The amount of value that boys and girls place on mathematics and science as well as the number of valued domains may influence gender differences in STEM achievement choices as well. The results are mixed on whether boys and girls differ in how much they value STEM domains, with many studies showing no gender differences in levels of STEM value (Eccles, 2009). However, there are gender differences in the number of valued domains, suggesting that women place high value on more domains (including non-STEM domains) than men do, which can lead to even high levels of STEM value being relatively less important for women (Eccles, 2007; Eccles, Barber, & Jozefowicz, 1999; Thoman, Arizaga, Smith, Story, & Soncuya, 2013). Additionally,

women, compared with men, tend to believe it is more important to make occupational sacrifices for the family and to have a job that helps people, which is one of the strongest predictors for women not pursuing STEM careers (Eccles, 2007). Men, however, are more likely to value making money and having a successful career. This difference may be especially crucial for talented girls, because they are caught between their beliefs in gender stereotypes on the one hand and their accomplishments in mathematics and science courses on the other (Eccles, 2007). Thus, high-achieving girls may shy away from enrolling in challenging STEM courses because of their belief in cultural stereotypes. Parents and other socializers, whose values are influenced by cultural stereotypes, may transmit these stereotyped beliefs to their adolescents.

Utility-Value Interventions

Recent studies have focused on understanding the particular role of utility value (UV) in achievement behaviors (Durik & Harackiewicz, 2007; Hulleman et al., 2008; Hulleman, Godes, Hendricks, & Harackiewicz, 2010; Hulleman & Harackiewicz, 2009; Kauffman & Husman, 2004; Shechter, Durik, Miyamoto, & Harackiewicz, 2011). For example, Hulleman et al. (2008) found that students' perceptions of utility value predicted achievement in both a college classroom and a high school sports camp. In another study, students who had higher utility value for their studies persisted longer and performed better than those who had lower levels (Vansteenkiste, Simons, Lens, Sheldon, & Deci, 2004).

On the basis of this correlational research, researchers have recently begun to manipulate utility value with interventions in the lab, classroom, and home (Acee & Weinstein, 2010; Durik & Harackiewicz, 2007; Harackiewicz et al., 2012; Hulleman et al., 2010; Hulleman & Harackiewicz, 2009). They have targeted utility value in particular because it is likely that perceptions of utility value can be changed with interventions. Attainment and intrinsic values are more intrinsic and therefore would be difficult for an outside entity to manipulate. Utility value, in contrast, should be amenable to change by an intervention. Studies have found that these utility value interventions cause an increase in interest and performance in the subject, including STEM topics (Durik & Harackiewicz, 2007; Hulleman et al., 2010; Hulleman & Harackiewicz, 2009; Shechter et al., 2011). Although these UV interventions have had positive effects on motivation, these effects have typically been moderated by past performance or expectations for success, which is consistent with expectancy-value theory (Nagengast et al., 2011; Trautwein et al., 2012). Individuals with high expectations for success responded most positively when told why a topic was relevant to their lives (e.g., Durik & Harackiewicz, 2007), whereas individuals with low expectations for success showed no positive response or responded negatively when given relevance (UV) information (for a review, see Durik, Hulleman, & Harackiewicz, 2013). These results suggest that it is critically important to consider the role of expectations and past performance in studies involving utility-value interventions.

Indirect Utility-Value Interventions

Based on the documented potential of UV information to promote motivation for many individuals and the associations between parents' values and their adolescent's values in correlational

research, we implemented a utility-value intervention aimed at parents (Harackiewicz et al., 2012). The ultimate goal of this intervention was to increase adolescents' STEM UV and STEM course-taking in high school. Previous research had not used randomized experiments to test the influence of parents on adolescents' utility value and achievement choices, but this study was able to evaluate the role of parents by randomly assigning them to an experimental UV intervention versus control condition. In the experimental condition, parents in an ongoing longitudinal study were given information about the relevance or usefulness (utility value) of mathematics and science for their adolescent. Parents in the control group received no information.

The results indicated that adolescents whose parents were in the intervention group took almost a semester more of mathematics and science classes during the last 2 years of high school than those whose parents were in the control group. These results indicated that parents can play a crucial role in increasing important adolescent achievement choices, such as advanced STEM course-taking. Although this intervention was effective for adolescents on average, it is important to consider the possibility that this intervention effect may vary as a function of gender and past performance, as has been observed in previous studies. It is also important to examine how this intervention worked to influence adolescents' course-taking.

The Current Study

This study goes beyond our previous evaluation of the utility-value intervention described above, to investigate for whom the intervention worked best and how it worked. The first research question asked whether gender and past performance (i.e., 9th-grade math and science grade point average) moderated the effects of the intervention. Previously, we found a main effect of the intervention on course-taking in the last 2 years of high school; later we coded past performance from high-school transcripts to use as a proxy for expectancies to test for an expectancy (prior performance) by value (intervention) interaction. Given the underrepresentation of women in many STEM fields (Halpern et al., 2007) and previously documented gender differences in expectancies and values in the STEM domain, we tested both gender and past performance as moderators of the intervention effect. Although in an earlier paper we reported that the intervention effect did not differ as a function of gender (Harackiewicz et al., 2012), we hypothesized that gender differences might emerge once we considered students' past performance. We therefore tested for an interaction among the intervention, gender, and past performance in STEM classes.

Using a mediation model, the second research question asked what mechanisms accounted for the effect of the intervention on students' course-taking (see Figure 1 for the theoretical model). We hypothesized that the intervention would lead to increased STEM UV for parents, which we assessed with questionnaires given to mothers of the adolescents. This increase in mothers' STEM UV was then predicted to be associated with an increase in adolescents' perceptions of parents' STEM values and adolescents' STEM UV. To provide the strongest test of mediation, we capitalized on the longitudinal design of the original study. The outcome variable was 12th-grade STEM course-taking. Mothers' perceived STEM UV, adolescents' perceptions of parents' STEM

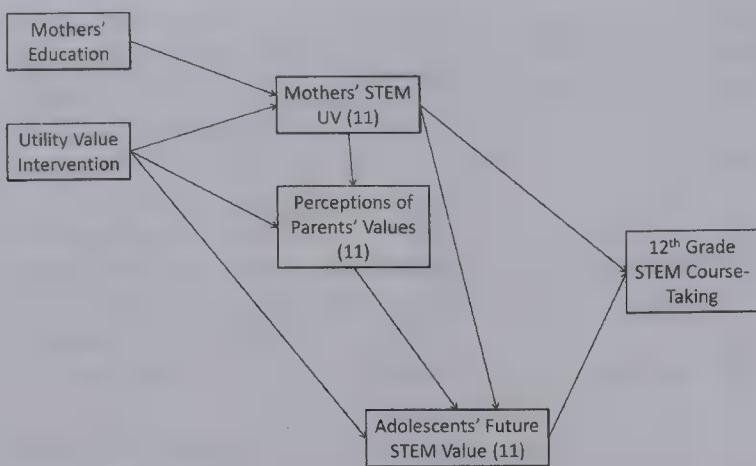


Figure 1. Theoretical model. STEM = science, technology, engineering, and mathematics; UV = utility value.

values, and adolescents' perceived STEM utility value were measured in the summer after 11th grade and therefore could be tested as mediators in the analyses of the effects of the intervention (which occurred during 10th and 11th grades) on 12th-grade STEM course-taking. These variables were predicted to mediate the effect of the intervention on 12th-grade STEM course-taking.

Method

Participants

The sample comprised families participating in the longitudinal Wisconsin Study of Families and Work (WSFW; for details on recruitment, see Hyde, Klein, Essex, & Clark, 1995). The current sample consisted of 188 adolescents (88 girls, 100 boys) and their parents who participated in a randomized experiment during high school (Harackiewicz et al., 2012). With regard to ethnicity, 90% of the adolescents were White (not of Hispanic origin), 2% were African American, 1% were Native American, and 7% were biracial or multiracial; this distribution is characteristic of the state of recruitment, in which 90% of the population is White (U.S. Census Bureau, 2006). At the time of data collection, participants attended 108 different high schools, increasing the generalizability of the findings. In 2010, the majority of adolescents (98%) had graduated from high school, and 94% reported plans to attend college or technical school. Average parents' years of education was 15.42 years ($SD = 1.92$) on a scale where 12 years is equivalent to high school graduation or GED completion.

Experimental Procedure

The intervention was implemented in October 2007 (10th grade) and again in January 2009 (11th grade). Families were followed through the teens' graduation from high school in June 2010. Families were randomly assigned to one of two experimental conditions and were blocked on gender of teen and mothers' educational level. Of these 188 families, 83 were in the experimental group and 105 were in the control group.

The intervention materials (two brochures and a website) were delivered exclusively to parents and focused on the usefulness of mathematics and science for adolescents. In particular, these ma-

terials explored potential connections between mathematics and science and current and future goals of adolescents (Harackiewicz et al., 2012). A first brochure, titled "Making Connections: Helping Your Teen Find Value in School," was sent to each household, addressed to the parents, in October of 10th grade. A second brochure, titled "Making Connections: Helping Your Teen with the Choices Ahead," was sent to each parent separately in January of 11th grade. This mailing included a letter giving them access to a dedicated, password-protected website called "Choices Ahead." Additionally, in the spring of 11th grade, parents in the experimental group were asked to complete an online questionnaire to evaluate the Choices Ahead website, which resulted in more parents visiting the website. A high percentage of parents (86%) reported using these resources, and a high percentage of adolescents (75%) reported exposure to this information. Parents in the control group did not receive any of these materials.

The 10th-grade brochure provided information about the importance or usefulness of mathematics and science in daily life and for various careers; it also provided parents with information about how to talk with adolescents about these issues. The 11th-grade brochure focused on these same themes but with different examples, and it gave greater emphasis to everyday activities (e.g., video games, cell phones) and preparation for college and careers. The 11th-grade brochure provided additional information for parents about communicating with their children about these issues and personalizing the relevance of mathematics and science for their 11th grader. The website featured clickable links to resources about STEM fields and careers. It also presented interviews with current college students who explained the usefulness of the mathematics and science courses that they had taken in high school. Parents were able to e-mail specific links from the site to their teens.

Measures

STEM courses taken in 12th grade and prior performance. Transcripts were obtained for 181 of the 188 students in the sample and came from 108 different high schools. Receipt of transcripts did not vary due to experimental condition or gender. The remaining sample of 181 families included 47 girls and 53 boys in the control group and 39 girls and 42 boys in the intervention group. For the outcome measure, we coded transcripts for the number of semesters of mathematics and science taken during 12th grade (12th-grade STEM course-taking). (Note that Harackiewicz et al. (2012) used number of mathematics and science courses taken in 11th and 12th grades as the outcome variable. Here we used just the number taken in 12th grade, so that a mediation model could be tested with mediators measured in 11th grade.)

For the measure of prior STEM performance, we created a standardized measure of ninth-grade STEM grade point average (GPA) by individually calculating each adolescent's GPA for mathematics and science courses taken in ninth grade on a GPA scale that ranged from 0 (F) to 4.0 (A/A+). The scale distinguished between grades by one third of a grade point (e.g., A = 4.0, A- = 3.67, B+ = 3.33). The final measure was a weighted, cumulative STEM GPA from ninth grade that took into account

the number of credits each course counted to weight the course grade.

Mother's STEM utility value, adolescent's perceptions of parents' values, and adolescents' future STEM value. Questionnaires given to mothers and adolescents in the summer after 11th grade included one measure from mothers (mothers' STEM UV for their adolescent) and two adolescent measures (perceptions of parents' STEM values and adolescents' STEM value). Response rates on the questionnaires were 83% for mothers and 77% for adolescents. All measures were based on items developed by Eccles and colleagues (e.g., Eccles & Wigfield, 2002; Eccles-Parsons et al., 1983). Mothers' STEM UV was measured with four items that asked about the mother's perceptions of the utility value of mathematics and science for her adolescent (e.g., *In general, how useful will [biology] be for your teen in the future?* $\alpha = .79$). This question was asked about four STEM topics: biology, mathematics, chemistry, and physics. Responses were on a scale from 1 (*not at all useful*) to 5 (*very useful*). Fathers also reported on STEM UV for their adolescent. However, the response rate for fathers at 11th grade was only 62%, creating substantial missing data. Therefore, we used only the variable from mothers.

For adolescents' perceptions of parents' values, adolescents rated how important their parents thought mathematics and science would be in their lives with two items (e.g., *My parents think math and science are important for my life*; $\alpha = .78$). Adolescents' perception of the value of mathematics and science for their future (future STEM value) was measured with four items that focused on the current and future value of mathematics and science for themselves (e.g., *Math and science are important for my future*; $\alpha = .79$). Adolescent measures were rated on a scale from 1 (*strongly disagree*) to 7 scale (*strongly agree*).

Parents' education. In the current sample ($N = 181$), mothers averaged 15.42 years of education ($SD = 2.10$), and fathers also averaged 15.42 years of education ($SD = 2.41$). A variable of parents' average years of education ($M = 15.42$, $SD = 1.92$) was created by averaging these two variables ($r = .44$). In this paper, we use mothers' education for analyses involving mother variables and parents' education for analyses not involving mothers' reports.

Overview of Analyses

We used multiple regression followed by structural equation modeling to analyze these data in two stages. First, multiple regression was used to investigate the direct effects of the predictors on 12th-grade STEM courses taken, which was the primary outcome variable. Second, a structural equation model was estimated based on the theoretical model (see Figure 1) to examine the relationships among the predictors, mediators (mothers' UV, perceptions of parents' values, and adolescents' future STEM value), and the outcome in a single model. In this model, we tested whether the total indirect effect of the predictors on the outcome through the mediators was significant (Preacher & Hayes, 2008). Cases with missing data were included by using full information maximum likelihood methods (Arbuckle, 1996).

There were seven predictors involving the intervention and the moderators of the intervention (*base predictors*): the intervention (coded as 1 for intervention group and -1 for control group), adolescent's gender (coded 1 for boys and -1 for girls), ninth-

grade STEM GPA (measured continuously and standardized), and two- and three-way interactions (the interaction of the intervention by adolescent's gender, the interaction between the intervention and ninth-grade STEM GPA, the interaction between adolescent's gender and ninth-grade STEM GPA, and the three-way interaction among the intervention, adolescent's gender, and ninth-grade STEM GPA). Finally, we included a term to test parental education.

Results

Zero-order correlations and descriptive statistics for all variables are shown in Table 1, separately by adolescent's gender.

Multiple Regression Model of Direct Effects on Course-Taking

To address the first research question, we regressed 12th-grade STEM courses taken on the base predictors and parents' education.¹ For 12th-grade STEM courses taken, there was one significant effect: the three-way interaction among the intervention, adolescent's gender, and ninth-grade STEM GPA ($z = -2.44$, $p < .05$, $\beta = -.18$).² In contrast to the main effect of the intervention reported by Harackiewicz et al. (2012), the pattern of the three-way interaction (see Figure 2) suggests that, when prior performance and gender are taken into consideration, the intervention increased course-taking for low-GPA boys ($\beta = .27$, $p < .05$) and high-GPA girls ($\beta = .22$, $p < .10$), whereas the intervention did not help low-GPA girls ($\beta = -.20$, trend toward a negative effect of the intervention) and had no effect on high-GPA boys ($\beta = -.04$). The graph of the three-way interaction in Figure 2, as for all interaction graphs in this paper, follows the convention of graphing high values at 1 SD above the mean of GPA and low values at 1 SD below the mean (Aiken & West, 1991).

Structural Equation Model

To address the second research question, we used structural equation modeling in Mplus to test whether the direct effect of the intervention (as moderated by gender and prior STEM performance) on 12th-grade STEM course-taking was mediated by indirect effects through the mediators. In the model (see Figure 1), we estimated paths from the base predictors (the intervention, gender, prior STEM performance, and their interactions) to mothers' STEM UV, perceptions of parents' values, and adolescents' future STEM value. To be consistent with previous analyses (Harackiewicz et al., 2012), we also included mothers' years of education as a predictor of mothers' STEM UV and of STEM course-taking. In accordance with the theoretical model, mothers' STEM

¹ The results remain the same if mothers' education is substituted for parents' education here. The three-way interaction is still the only significant predictor ($z = -2.39$, $p < .05$, $\beta = -.18$).

² These regression analyses were repeated with STEM course-taking in 11th and 12th grades as the outcome measure, the one used in the Harackiewicz et al. (2012) paper. The results were the same, that is the three-way interaction among intervention, gender, and prior performance significantly predicted 11th- plus 12th-grade STEM course-taking. We report results in detail here only for the 12th-grade course-taking outcome, to preserve the temporal sequence for mediation analyses.

Table 1

Zero-Order Correlations and Descriptive Statistics for Major Variables by Gender

Variable	1	2	3	4	5	6	7
1. Ninth-grade STEM GPA	—	0.34**	0.36**	0.26*	0.26*	0.24*	0.18
2. Mothers' STEM UV	0.21	—	0.54**	0.39**	0.27*	0.30*	0.27*
3. Adolescents' future STEM UV	0.40**	0.52**	—	0.55**	0.34**	0.28*	0.15
4. Perceptions of parents' values	0.37**	0.39**	0.61**	—	0.15	0.25*	0.16
5. STEM courses (12th grade)	0.16	0.34**	0.36**	0.18	—	0.11	0.04
6. Parents' education	0.42**	0.15	0.10	0.17	0.26*	—	0.79**
7. Mothers' education	0.42**	0.27*	0.26*	0.34*	0.21	0.86**	—
Girls, <i>M</i> (<i>SD</i>)	3.15 (0.84)	4.08 (0.79)	5.23 (1.43)	5.75 (1.06)	3.77 (1.71)	15.35 (2.09)	15.41 (2.33)
Boys, <i>M</i> (<i>SD</i>)	2.92 (0.88)	4.11 (0.81)	5.03 (1.63)	5.62 (1.26)	3.45 (1.85)	15.48 (1.76)	15.43 (1.88)

Note. Correlations above the diagonal are for boys. Correlations below the diagonal are for girls. There were no mean differences due to gender. STEM = science, technology, engineering, and mathematics; GPA = grade point average; UV = utility value.

p* < .05. *p* < .01.

UV was an additional predictor of perceptions of parents' values and adolescents' future STEM value. Furthermore, perception of parents' values was a predictor of adolescents' future STEM value. Additionally, paths were estimated from the base predictors, mothers' STEM UV, and adolescents' future STEM value to 12th-grade STEM courses taken. Thus, by examining the indirect effects of the base predictors through the mediators to STEM course-taking, this model tested whether the intervention, as moderated by GPA and adolescent's gender, influenced STEM course-taking through mothers' STEM UV, adolescents' perceptions of parents' values, and adolescents' future STEM value. Because this is a saturated model, it does not allow for a meaningful test of model fit.

Overall, the model accounted for 16.8% of the variance in 12th-grade STEM course-taking, 13.9% of the variance in mothers' STEM UV, 26.7% of the variance in perceptions of parents' values, and 50.8% of the variance in adolescents' future STEM value. See Figure 3 for the path models showing these results.

Effects on mothers' STEM UV. The base predictors and years of mothers' education were used to predict mothers' STEM

UV. There was a nearly significant effect of ninth-grade STEM GPA ($z = 1.94, p = .06, \beta = .17$) showing a trend for mothers to perceive more STEM utility value when their adolescent had a higher ninth-grade STEM GPA. In addition, the predicted three-way interaction among the intervention, adolescent's gender, and ninth-grade STEM GPA was significant ($z = -1.96, p = .05, \beta = -.16$); it is graphed in Panel A of Figure 4. The pattern of this interaction effect is similar to the one for the course-taking outcome in the multiple regression analysis (see Figure 2). Finally, mothers' education was a significant predictor of mothers' STEM UV ($z = 2.32, p < .05, \beta = .20$), such that mothers with more years of education showed higher levels of STEM UV.³

Effects on perceptions of parents' values. The base predictors and mothers' STEM UV were used to predict adolescents' perceptions of parents' values. There were significant effects of ninth-grade STEM GPA ($z = 2.64, p < .05, \beta = .23$), such that parents were perceived as seeing the value of STEM course-taking more when the adolescent had a higher STEM GPA. The two-way interaction between adolescent's gender and the intervention was significant ($z = 2.41, p < .05, \beta = .19$), suggesting that the intervention increased boys' perceptions of parents' values and decreased girls' perceptions of parents' values; however, this two-way interaction was qualified by the three-way interaction among the intervention, adolescent's gender, and ninth-grade STEM GPA, which was nearly significant ($z = -1.89, p = .06, \beta = -.17$). The pattern of the interaction is similar to the one for course-taking; in particular, the intervention appeared to decrease low-GPA girls' perceptions of their parents' values for them (see Figure 3, Panel A, and Figure 4, Panel B). That is, low-GPA girls in the intervention group perceived a lack of support for STEM from their parents. Finally, mothers' STEM UV was a significant predictor of adolescents' perceptions of parents' values ($z = 3.70, p < .01, \beta = .29$), such that mothers with higher levels of STEM UV tended to have adolescents with higher levels of perceptions of parents' values.

Effects on adolescents' future STEM value. The base predictors, mothers' STEM UV, and perceptions of parents' values

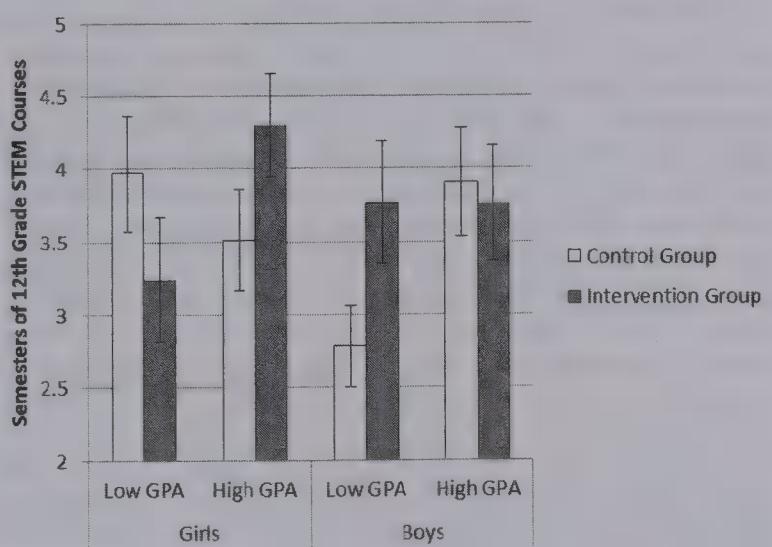
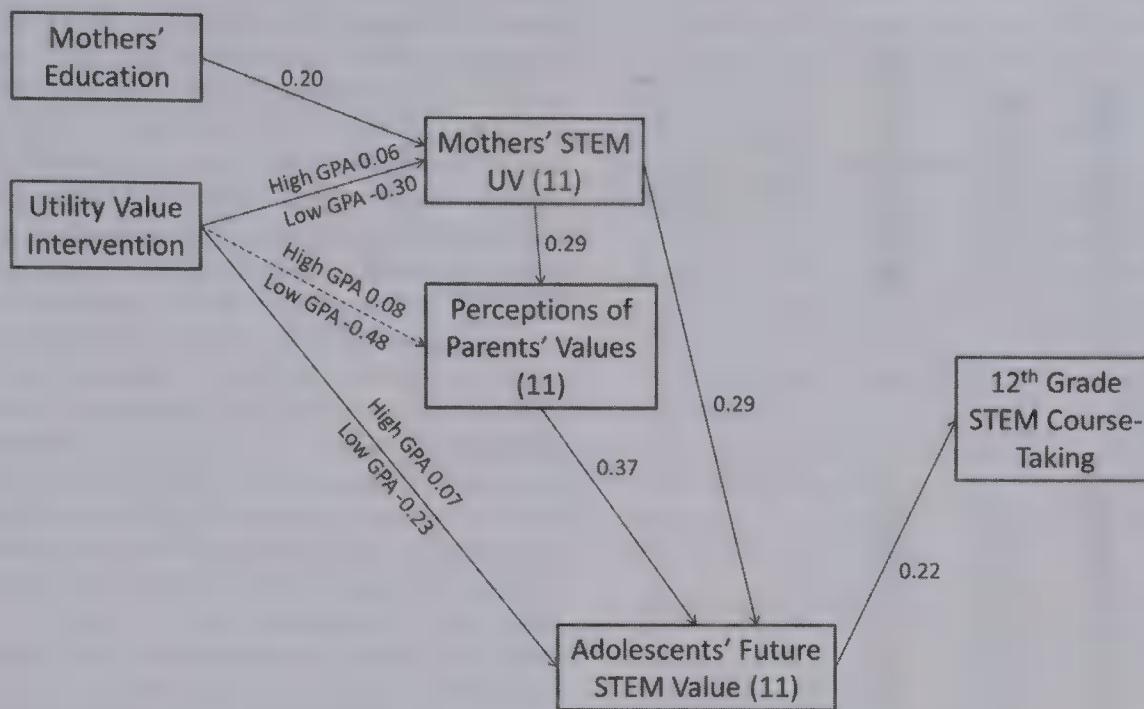


Figure 2. Direct effects of the intervention, adolescent's gender, and ninth-grade STEM GPA on STEM course-taking (12th grade). Predicted values were generated for high (1 *SD* above the mean) and low (-1 *SD*) ninth-grade STEM GPA from the multiple regression models. Error bars represent ± 1 *SEM*. STEM = science, technology, engineering, and mathematics; GPA = grade point average; SEM = standard error of the mean.

³ The model was also tested using parents' education instead of mothers' education, and the results for the overall model did not change; however, parents' education was a nonsignificant predictor of mothers' STEM UV ($z = 1.77, p > .05, \beta = .15$).

A. Intervention Effects for Girls



B. Intervention Effects for Boys

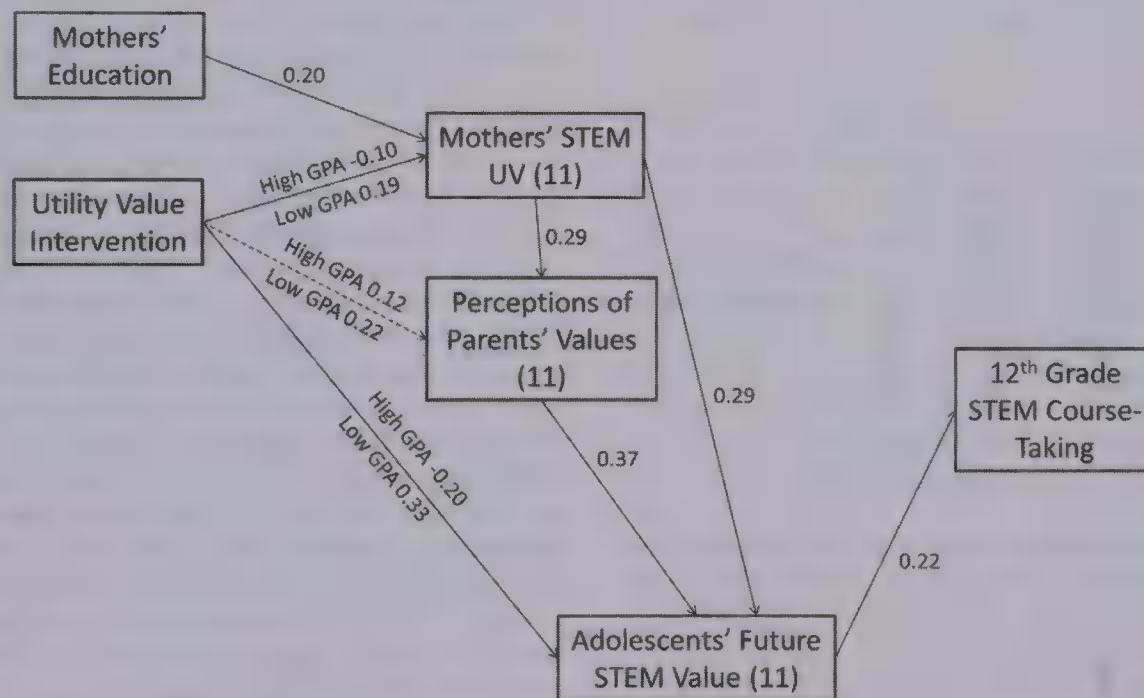


Figure 3. Empirical path model. Only significant paths are shown. The effect of the intervention on STEM course-taking in 12th grade differs by gender and ninth-grade STEM GPA and is mediated by mother's utility value (UV), perceptions of parents' values, and adolescents' future STEM value. The different intervention effects are shown (A) for girls and (B) for boys. Dashed line indicates $p = .06$ (the three-way interaction involving the intervention to perceptions of parents' values). STEM = science, technology, engineering, and mathematics; GPA = grade point average.

were used to predict adolescents' future STEM value (see Figure 3). The three-way interaction among the intervention, adolescent's gender, and ninth-grade STEM GPA was significant, as predicted ($z = -2.85$, $p < .05$, $\beta = -.21$). The three-way interaction is shown in Panel C of Figure 4; the pattern of the interaction is also

similar to the one for STEM course-taking and suggests that the intervention increased adolescents' future STEM value particularly for low-GPA boys. The effect of mothers' STEM UV was significant ($z = 4.35$, $p < .01$, $\beta = .29$); higher levels of mothers' STEM UV predicted higher levels of adolescents' future STEM

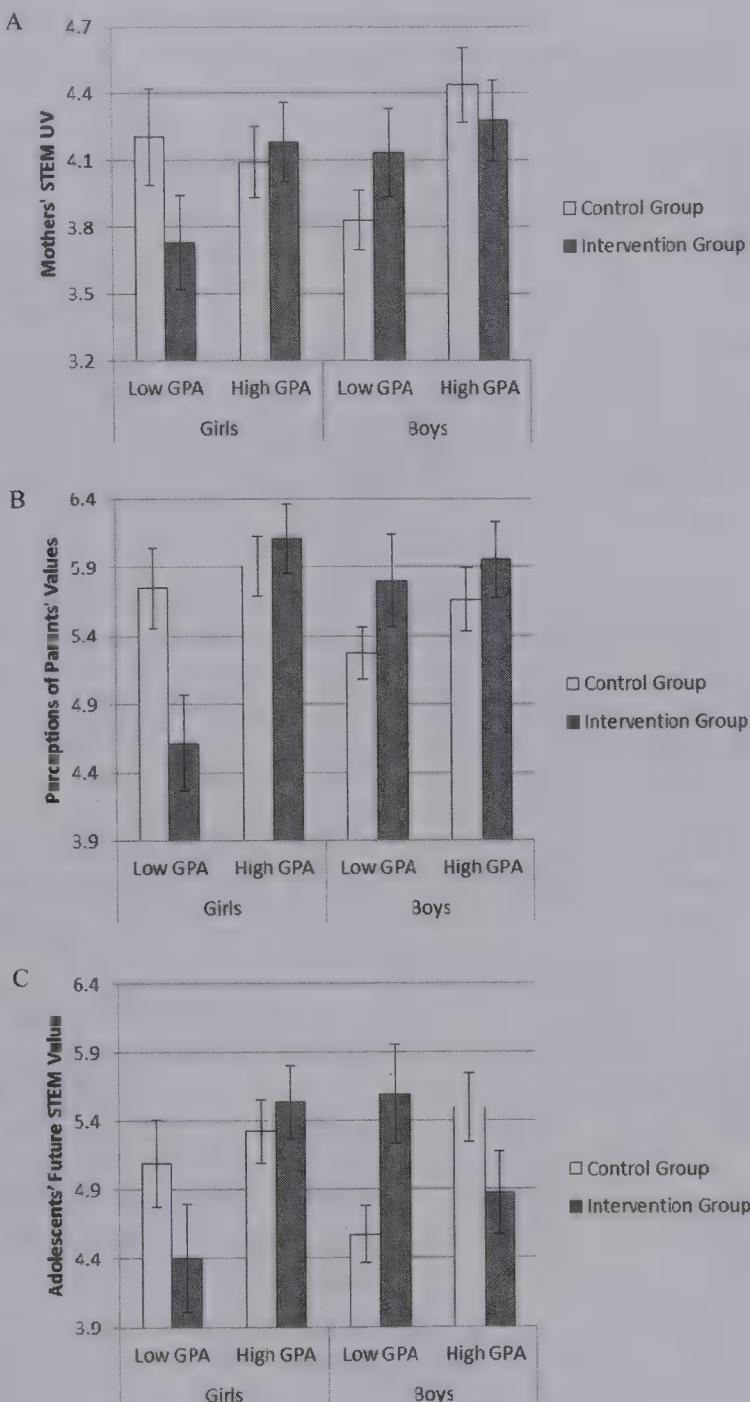


Figure 4. Direct effects of the intervention, child gender, and ninth-grade STEM GPA on the hypothesized mediators: (A) mothers' STEM UV, (B) perceptions of parents' values, and (C) adolescents' STEM utility value. Predicted values were generated for high (1 SD above the mean) and low (-1 SD) ninth-grade STEM GPA from the multiple regression models. For (A) the range of possible values is 1 to 5. For (B and C) the range of possible values is 1 to 7. Error bars represent ± 1 SEM. STEM = science, technology, engineering, and mathematics; GPA = grade point average; UV = utility value; SEM = standard error of the mean.

value. Perceptions of parents' values was a significant predictor as well ($z = 5.36, p < .01, \beta = .37$); higher levels of perceptions of parents' values predicted higher levels of adolescents' future STEM value.

Effects on 12th-grade STEM course-taking. The base predictors, mothers' STEM UV, adolescents' future STEM value, and mothers' years of education were used to predict 12th-grade STEM course-taking.⁴ There was a significant effect of adoles-

cents' future STEM UV ($z = 2.18, p < .05, \beta = .22$). Higher levels of adolescents' future STEM value predicted more 12th-grade STEM courses taken, for both boys and girls.

Indirect effects and mediation. In the structural equation model, we hypothesized that the base predictors (specifically the three-way interaction) would influence 12th-grade STEM course-taking through the mediators, so the direct effects of the base predictors to 12th-grade STEM course-taking shown in the direct effects model should be reduced in a model containing the mediators; additionally, there should be significant indirect effects of the three-way interaction through the mediators to 12th-grade STEM course-taking. We hypothesized that the mediation would work in a specific way; that is, the three-way interaction should predict mothers' STEM UV, perceptions of parents' values, and adolescents' future STEM UV. Mothers' STEM UV should predict perceptions of parents' values and adolescents' future STEM UV, and perceptions of parents' values should predict adolescents' future STEM UV. Additionally, we specified that mothers' STEM UV and adolescents' future STEM value would predict STEM course-taking. Using procedures described by Preacher and Hayes (2008), we tested the total indirect effect of the intervention through the three mediators, as well as the specific indirect effect of mothers' STEM UV through adolescents' perceptions of parents' values and adolescents' future STEM UV.

Therefore, two indirect pathways were tested in order to test for the indirect effect of the three-way interaction on 12th-grade STEM course-taking as well as the indirect effect of mothers' STEM UV on 12th-grade STEM course-taking. For the first, we tested whether the three-way interaction had a significant total indirect effect on 12th-grade STEM course-taking through the three mediators and found support for this hypothesis ($z = -2.40, p < .05$). Therefore, the intervention, as moderated by adolescent's gender and ninth-grade STEM GPA, had a significant total indirect effect on course-taking through the mediating variables: mothers' STEM UV, perceptions of parents' values, and adolescents' future STEM value. Additionally, the model with the mediators reduced the direct effects of the predicted three-way interaction on 12th-grade STEM course-taking (direct effect, $\beta = -0.18, p < .05$; with mediators in the model, $\beta = -0.09, ns$).

For the second, we tested for the specific indirect effect of mothers' STEM UV to 12th-grade STEM course-taking through perceptions of parents' values and adolescents' future STEM value. Results indicated a significant specific indirect effect ($z = 2.06, p < .05$). This indicated that mothers' STEM UV had a significant specific indirect effect on 12th-grade STEM course-taking through perceptions of parents' values and adolescents' future STEM value.

Discussion

To address concerns about low rates of adolescents taking advanced STEM courses in high school in the United States, we implemented an intervention, based in expectancy-value theory, with parents of adolescents (Harackiewicz et al., 2012). In the results reported here, we examined whether the intervention was differentially effective for girls compared with boys in the context

⁴ The analyses were repeated using parents' education instead of mothers' education. Findings remained unchanged.

of past performance and what factors mediated the effects of the intervention on course-taking. In response to the first research question, the results from multiple regression analysis indicated that the intervention increased STEM course-taking in 12th grade for girls who had done well in ninth-grade STEM courses (high GPA) and for boys who had not done well (low GPA). However, the intervention did not increase course-taking for low-GPA girls (trending toward a negative effect), and it had no effect for high-GPA boys. The absence of an effect for high-GPA boys is most likely due to a ceiling effect on the measure of number of STEM courses taken in 12th grade.

In regard to the second research question, mediation analyses suggested that these intervention effects (specifically the three-way interaction among the intervention, gender, and prior STEM performance) occurred through changes in both mother and adolescent variables. The intervention was targeted exclusively at parents, so we predicted and found that the intervention increased mothers' STEM utility value for their adolescents. The intervention also led adolescents to perceive higher levels of parental STEM values and increased adolescents' future STEM value, and the changes in mothers' STEM utility value contributed to these changes in adolescent variables. Overall, the effect of the intervention on high-school STEM course-taking was mediated by the effects of the intervention on mothers' STEM utility value and adolescents' STEM utility value. This suggests that parents' utility value does indeed influence adolescents' utility value and achievement behavior.

Considerable support for Eccles's expectancy-value theory has been amassed through correlational and longitudinal research, but experimental support has been lacking. One strength of an experimental approach to this theory is that researchers can assess the causal effect of task values on achievement motivation and behavior. In particular, when studying families, an association has been shown between parents' beliefs and their children's beliefs and achievement-related behaviors (e.g., Chhin, Bleeker, & Jacobs, 2008), but the direction of the effect has been unclear. To explore whether parents' values could influence adolescents' values, we experimentally manipulated parents' utility value through a randomized intervention to assess the causal impact of parents' beliefs on their children's beliefs and behaviors (Harackiewicz et al., 2012). Although the original study showed that an increase in adolescents' STEM course-taking over the final 2 years of high school occurred as a result of this intervention, mediation analyses of this effect were not conducted.

Processes Underlying Intervention Effects

In the current paper, we examined the hypothesis that the intervention worked by changing parents' and adolescents' STEM utility value. We found support for this hypothesis. In our previous paper (Harackiewicz et al., 2012), the results indicated that the intervention affected mothers' STEM utility value, which provides crucial support that this utility value intervention for parents had its intended effect. In the current analyses, this increase in mothers' STEM utility value was related to an increase in adolescents' perceptions of how much their parents valued STEM for them and also adolescents' future STEM value. Thus, both mothers and adolescents had increased perceptions of STEM value due to the intervention. Because the intervention was targeted exclusively at

parents, it is reasonable to conclude that adolescents were influenced by their parents.

Two paths in Figure 3 warrant additional discussion. First, the direct path (specifically the three-way interaction among the intervention, gender, and prior STEM performance) from the intervention to adolescents' perceptions of their parents' values was significant, above and beyond the indirect path through mothers' STEM utility value. That is, the intervention appeared to have some effect on adolescents' perceptions beyond the effect it had on mothers' STEM UV for them. This might involve a process such as a mother sharing the intervention website with her adolescent while not expressing her beliefs in the value of STEM. Second, the direct path from mothers' STEM UV to adolescents' future STEM value was significant, beyond the indirect effect through adolescents' perceptions of their parents' values. This effect might involve some changes in mothers' behavior that are not consciously perceived by the adolescent but that nonetheless have an effect.

Moderation by Gender and Prior Performance

In this paper we also considered whether the intervention, which had an overall positive main effect on course-taking, might be differentially effective based on the adolescent's gender and prior STEM performance. The results indicated that, in fact, adolescents' prior STEM grades moderated the effect of the utility value intervention differently for girls and boys. The intervention had positive effects on STEM course-taking for low-GPA boys and high-GPA girls, but it had no effect (trending toward a negative effect) for low-GPA girls and had no effect for high-GPA boys.

Why were low-GPA girls not helped by the intervention when low-GPA boys were helped by it? The measure of prior performance, ninth-grade STEM GPA, should be linked tightly to both mothers' and adolescents' expectations for future success in STEM and has been used as a proxy for expectations in previous utility intervention research (Hulleman et al., 2010). Yet, research shows that parents are more likely to have inflated expectancies for success for boys in this domain in comparison to girls (Eccles, Jacobs, & Harold, 1990; Gunderson, Ramirez, Levine, & Beilock, 2012; Jacobs, Davis-Kean, Bleeker, Eccles, & Malanchuk, 2005; Yee & Eccles, 1988). Thus, parents may assess all boys as capable of success in STEM, even if they have had low grades in school. Therefore, even low-GPA boys may benefit from a utility-value intervention targeted at parents, because parents will still deem them capable of succeeding. Boys with higher prior STEM achievement did not benefit from the intervention, probably due to a ceiling effect in the number of semesters of mathematics and science taken during 12th grade. That is, their STEM course-taking was constrained by factors such as the number of class periods in the day and requirements that they take non-STEM courses. Positive effects of the intervention for high-GPA boys might be revealed in situations with fewer constraints (e.g., in college).

For girls, low STEM GPA may create low expectations for success—both for the girl and her mother—that negate the beneficial effects of the UV intervention; even if parents see the value of STEM, their low expectations for success for their low-GPA daughters mean that parents have low STEM aspirations for them, rendering the utility value of STEM irrelevant. These effects are consistent with the predicted effects in Eccles's expectancy-value theory. Moreover, they are consistent with past research showing

that UV interventions are less effective for those with low expectations for success (e.g., Durik & Harackiewicz, 2007).

In addition, girls and their mothers observe the unbalanced gender composition of many adult occupations (Ridgeway, 2011), which may contribute to the findings. Whereas girls with a high STEM GPA may aspire to traditionally masculine careers requiring substantial mathematics and science and be responsive to the intervention, girls with a low STEM GPA may see no reason to consider such aspirations and, simultaneously, may be drawn to traditionally feminine careers such as child-care worker (95% female, Bureau of Labor Statistics, 2011) or elementary- or middle-school teacher (82% female), which appear to require little mathematics and science (Beilock, Gunderson, Ramirez, & Levine, 2010). Moreover, if parents share these beliefs, they may not encourage their daughters to pursue STEM careers. This interpretation is supported by the relatively low level of parental valuing of STEM that low-GPA girls reported (see Figure 4, Panel B). On balance, then, girls with low prior STEM performance may have little interest in STEM courses and careers and receive little encouragement from parents, despite the intervention, while simultaneously experiencing a strong pull toward traditionally female careers that appear to require little mathematics and science and where they feel that they "belong" (Thoman et al., 2013).

It will be important for future interventions to take into account the role of expectancies in designing utility-value interventions that will be successful for all students. Recent research has shown that, although the interactive effects of expectancy and value are mixed, this interaction does occur in some studies (Nagengast et al., 2011; Trautwein et al., 2012). This intervention was in the STEM domain, so it is likely that both parents' and adolescents' expectancies would be affected not only by prior achievement but also by the adolescent's gender. Future interventions may be strengthened by the inclusion of information that enhances not only perceptions of utility value but also expectations for success.

Limitations and Directions for Future Research

Several limitations should be kept in mind when interpreting these results. First, the sample was representative of the state of Wisconsin but not racially diverse, so future research should extend these findings to more diverse groups. Previous studies have shown that the effects of utility-value interventions are consistent across racial groups (Hulleman & Harackiewicz, 2009), suggesting that our results would extend to more diverse contexts. Additionally, although the sample size was sufficiently large to have the power to detect the intervention effects, future studies would benefit from scaling up the intervention to larger samples.

Second, although the utility-value intervention affected mothers' and adolescents' perceptions of utility value and adolescents' course-taking behavior, we do not have measures of the precise interpersonal processes by which these increases in mothers' utility value changed adolescents' attitudes. Correlational research has shown that these effects may be explained through a variety of parental behaviors, such as modeling, encouragement, and coactivity (e.g., Simpkins et al., 2012). Future studies could also assess these behaviors to understand how parents' perceptions of utility value result in behavioral change that affects their children. It is likely that parents use a variety of methods and behaviors to influence their children, so understanding which behaviors are

most effective will make an important contribution to future research. We believe that future studies may also benefit from using measures of adolescents' perceptions of their parents' values as we did here, because that measure can capture the effect of a variety of parental behaviors.

Third, this utility-value intervention (and much of the correlational research based on expectancy-value theory) was conducted within a specific domain, STEM. Therefore, we cannot assume that these intervention results would generalize to non-STEM domains, and future research should extend these findings to other domains. Previous research has shown that the relationships between utility value and achievement behavior do extend to non-STEM domains (e.g., Jodl et al., 2001), so the intervention effects should also generalize, but this will need to be tested in future studies.

Finally, although the utility-value intervention had effects that differed due to gender and prior achievement, it is important to recognize that, on average, this intervention had substantial positive effects on STEM course-taking (Harackiewicz et al., 2012). Future studies may modify this intervention to make it more effective, but it had generally positive effects on a key educational outcome needed to enhance STEM preparation. Therefore, we can recommend this intervention as having positive effects and also recommend taking into account expectancies for success to make it more effective in future research.

Implications

Several implications flow from these results. The findings indicate that parents are a resource—a largely untapped one—that may be used to enhance STEM motivation of adolescents. There is room to increase how much parents value STEM for their adolescents, and changes in parents' utility value can affect adolescents' beliefs and behavior. Therefore, parents—in addition to teachers and curriculum—may be used to increase students' STEM preparation and motivation. Future utility-value interventions should also attend to issues of expectations for success, particularly in regard to gender gaps in STEM.

References

- Acee, T. W., & Weinstein, C. (2010). Effects of a value-reappraisal intervention on statistics students' motivation and performance. *Journal of Experimental Education*, 78, 487–512. doi:10.1080/00220970903352753
- Aiken, L. S., & West, S. G. (1991). *Multiple regression: Testing and interpreting interactions*. Thousand Oaks, CA: Sage.
- Arbuckle, J. L. (1996). Full information estimation in the presence of incomplete data. In G. A. Marcoulides & R. E. Schumaker (Eds.), *Advanced structural equation modeling: Issues and techniques* (pp. 243–277). Mahwah, NJ: Erlbaum.
- Beilock, S. L., Gunderson, E. A., Ramirez, G., & Levine, S. C. (2010). Female teachers' math anxiety affects girls' math achievement. *Proceedings of the National Academy of Sciences, USA*, 107, 1860–1863. doi:10.1073/pnas.0910967107
- Bleeker, M. M., & Jacobs, J. E. (2004). Achievement in math and science: Do mothers' beliefs matter 12 years later? *Journal of Educational Psychology*, 96, 97–109. doi:10.1037/0022-0663.96.1.97
- Bureau of Labor Statistics. (2011). *Women at work*. Retrieved from <http://www.bls.gov/spotlight/2011/women/>
- Chhin, C. S., Bleeker, M. M., & Jacobs, J. E. (2008). Gender-typed occupational choices: The long-term impact of parents' beliefs and

expectations. In H. M. G. Watt & J. S. Eccles (Eds.), *Gender and occupational outcomes: Longitudinal assessments of individual, social, and cultural influences* (pp. 215–234). doi:10.1037/11706-008

Durik, A. M., & Harackiewicz, J. M. (2007). Different strokes for different folks: How individual interest moderates the effects of situational factors on task interest. *Journal of Educational Psychology*, 99, 597–610. doi:10.1037/0022-0663.99.3.597

Durik, A. M., Hulleman, C. S., & Harackiewicz, J. M. (2013). One size fits some: Instructional enhancements to promote interest don't work the same for everyone. In K. A. Renninger & M. Nieswandt (Eds.), *Interest, the self, and K-16 mathematics and science learning*. Washington, DC: American Educational Research Association.

Eccles, J. S. (2007). Where are all the women? Gender differences in participation in physical science and engineering. In S. J. Ceci & W. M. Williams (Eds.), *Why aren't more women in science: Top researchers debate the evidence* (pp. 199–210). doi:10.1037/11546-016

Eccles, J. (2009). Who am I and what am I going to do with my life? Personal and collective identities as motivators of action. *Educational Psychologist*, 44, 78–89. doi:10.1080/00461520902832368

Eccles, J. S., Barber, B., & Jozefowicz, D. (1999). Linking gender to educational, occupational, and recreational choices: Applying the Eccles et al. model of achievement-related choices. In W. B. Swann, J. H. Langlois, & L. A. Gilbert (Eds.), *Sexism and stereotypes in modern society: The gender science of Janet Taylor Spence* (pp. 153–192). doi:10.1037/10277-007

Eccles, J. S., Barber, B. L., Updegraff, K., & O'Brien, K. M. (1998). An expectancy-value model of achievement choices: The role of ability self-concepts, perceived task utility and interest in predicting activity choice and course enrollment. In L. Hoffman, A. Krapp, K. A. Renninger, & J. Baumert (Eds.), *Interest and learning: Proceedings of the Second Conference on Interest and Gender* (pp. 267–280). Kiel, Germany: IPN.

Eccles, J. S., Jacobs, J. E., & Harold, R. D. (1990). Gender role stereotypes, expectancy effects, and parents' socialization of gender differences. *Journal of Social Issues*, 46, 183–201. doi:10.1111/j.1540-4560.1990.tb01929.x

Eccles, J. S., & Wigfield, A. (2002). Motivational beliefs, values, and goals. *Annual Review of Psychology*, 53, 109–132. doi:10.1146/annurev.psych.53.100901.135153

Eccles, J., Wigfield, A., Harold, R. D., & Blumenfeld, P. (1993). Age and gender differences in children's self- and task perceptions during elementary school. *Child Development*, 64, 830–847. doi:10.2307/1131221

Eccles-Parsons, J., Adler, T. F., Futterman, R., Goff, S. B., Kaczala, C. M., Meece, J. L., & Midgley, C. (1983). Expectancies, values, and academic behaviors. In J. T. Spence (Ed.), *Achievement and achievement motivation* (pp. 75–146). San Francisco, CA: Freeman.

Frome, P. M., & Eccles, J. S. (1998). Parents' influence on children's achievement-related perceptions. *Journal of Personality and Social Psychology*, 74, 435–452. doi:10.1037/0022-3514.74.2.435

Grolnick, W. S., & Ryan, R. M. (1989). Parent styles associated with children's self-regulation and competence in school. *Journal of Educational Psychology*, 81, 143–154. doi:10.1037/0022-0663.81.2.143

Grolnick, W. S., Ryan, R. M., & Deci, E. L. (1991). Inner resources for school achievement: Motivational mediators of children's perceptions of their parents. *Journal of Educational Psychology*, 83, 508–517. doi:10.1037/0022-0663.83.4.508

Gunderson, E. A., Ramirez, G., Levine, S. C., & Beilock, S. L. (2012). The role of parents and teachers in the development of gender-related math attitudes. *Sex Roles*, 66, 153–166. doi:10.1007/s11199-011-9996-2

Halpern, D. F., Benbow, C. P., Geary, D. C., Gur, R., Hyde, J. S., & Gernsbacher, M. A. (2007). The science of sex differences in science and mathematics. *Psychological Science in the Public Interest*, 8, 1–51. doi:10.1111/j.1529-1006.2007.00032.x

Harackiewicz, J. M., Canning, E. A., Tibbetts, Y., Giffen, C. J., Blair, S. S., Rouse, D. I., & Hyde, J. S. (2014). Closing the social class achievement gap for first-generation students in undergraduate biology. *Journal of Educational Psychology*, 106, 375–389.

Harackiewicz, J. M., Rozek, C. S., Hulleman, C. S., & Hyde, J. S. (2012). Helping parents to motivate adolescents in mathematics and science: An experimental test of a utility-value intervention. *Psychological Science*, 23, 899–906. doi:10.1177/0956797611435530

Hulleman, C. S., Durik, A. M., Schweigert, S. B., & Harackiewicz, J. M. (2008). Task values, achievement goals, and interest: An integrative analysis. *Journal of Educational Psychology*, 100, 398–416. doi:10.1037/0022-0663.100.2.398

Hulleman, C. S., Godes, O., Hendricks, B. L., & Harackiewicz, J. M. (2010). Enhancing interest and performance with a utility value intervention. *Journal of Educational Psychology*, 102, 880–895. doi:10.1037/a0019506

Hulleman, C. S., & Harackiewicz, J. M. (2009, December 4). Promoting interest and performance in high school science classes. *Science*, 326, 1410–1412. doi:10.1126/science.1177067

Hyde, J. S., Klein, M. H., Essex, M. J., & Clark, R. (1995). Maternity leave and women's mental health. *Psychology of Women Quarterly*, 19, 257–285. doi:10.1111/j.1471-6402.1995.tb00291.x

Jacobs, J. E., Davis-Kean, P., Bleeker, M., Eccles, J. S., & Malanchuk, O. (2005). "I can, but I don't want to": The impact of parents, interests, and activities on gender differences in math. In A. M. Gallagher & J. C. Kaufman (Eds.), *Gender differences in mathematics: An integrative psychological approach* (pp. 246–263). New York, NY: Cambridge University Press.

Jacobs, J. E., & Eccles, J. S. (1992). The impact of mothers' gender-role stereotypic beliefs on mothers' and children's ability perceptions. *Journal of Personality and Social Psychology*, 63, 932–944. doi:10.1037/0022-3514.63.6.932

Jodl, K. M., Michael, A., Malanchuk, O., Eccles, J. S., & Sameroff, A. (2001). Parents' roles in shaping early adolescents' occupational aspirations. *Child Development*, 72, 1247–1266. doi:10.1111/1467-8624.00345

Kauffman, D. F., & Husman, J. (2004). Effects of time perspective on student motivation: Introduction to a special issue. *Educational Psychology Review*, 16, 1–7. doi:10.1023/B:EDPR.0000012342.37854.58

Miyake, A., Kost-Smith, L. E., Finkelstein, N. D., Pollock, S. J., Cohen, G. L., & Ito, T. A. (2010, November 26). Reducing the gender achievement gap in college science: A classroom study of values affirmation. *Science*, 330, 1234–1237. doi:10.1126/science.1195996

Nagengast, B., Marsh, H. W., Scalas, L. F., Xu, M. K., Hau, K.-T., & Trautwein, U. (2011). Who took the "X" out of expectancy-value theory? A psychological mystery, a substantive-methodological synergy, and a cross-national generalization. *Psychological Science*, 22, 1058–1066. doi:10.1177/0956797611415540

National Science Foundation. (2012). *Science and engineering indicators 2012*. Retrieved from www.nsf.gov/statistics/seind12/

Paulson, S. E., & Sputa, C. L. (1996). Patterns of parenting during adolescence: Perceptions of adolescents and parents. *Adolescence*, 31, 369–381.

Preacher, K. J., & Hayes, A. F. (2008). Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. *Behavior Research Methods*, 40, 879–891. doi:10.3758/BRM.40.3.879

Ratelle, C. F., Larose, S., Guay, F., & Senécal, C. (2005). Perceptions of parental involvement and support as predictors of college students' persistence in a science curriculum. *Journal of Family Psychology*, 19, 286–293. doi:10.1037/0893-3200.19.2.286

Ridgeway, C. L. (2011). *Framed by gender: How gender inequality persists in the modern world*. New York, NY: Oxford University Press.

Riegle-Crumb, C., & King, B. (2010). Questioning a White male advantage in STEM: Examining disparities in college major by gender and race/ethnicity. *Educational Researcher*, 39, 656–664. doi:10.3102/0013189X10391657

Shechter, O. G., Durik, A. M., Miyamoto, Y., & Harackiewicz, J. M. (2011). The role of utility value in achievement behavior: The importance of culture. *Personality and Social Psychology Bulletin*, 37, 303–317. doi:10.1177/0146167210396380

Simpkins, S. D., Davis-Kean, P. E., & Eccles, J. S. (2006). Math and science motivation: A longitudinal examination of the links between choices and beliefs. *Developmental Psychology*, 42, 70–83. doi:10.1037/0012-1649.42.1.70

Simpkins, S. D., Fredricks, J. A., & Eccles, J. S. (2012). Charting the Eccles' expectancy-value model from mothers' beliefs in childhood to youths' activities in adolescence. *Developmental Psychology*, 48, 1019–1032. doi:10.1037/a0027468

Spera, C. (2005). A review of the relationship among parenting practices, parenting styles, and adolescent school achievement. *Educational Psychology Review*, 17, 125–146. doi:10.1007/s10648-005-3950-1

Spera, C. (2006). Adolescents' perceptions of parental goals, practices, and styles in relation to their motivation and achievement. *Journal of Early Adolescence*, 26, 456–490. doi:10.1177/0272431606291940

Thoman, D. B., Arizaga, J. A., Smith, J. L., Story, T. S., & Soncuya, G. (2013). The grass is greener in non-science, technology, engineering, and math classes: Examining the role of competing belonging to undergraduate women's vulnerability to being pulled away from science. *Psychology of Women Quarterly*. Advance online publication. doi:10.1177/0361684313499899

Trautwein, U., Marsh, H. W., Nagengast, B., Lüdtke, O., Nagy, G., & Jonkmann, K. (2012). Probing for the multiplicative term in modern expectancy-value theory: A latent interaction modeling study. *Journal of Educational Psychology*, 104, 763–777. doi:10.1037/a0027470

Updegraff, K. A., Eccles, J. S., Barber, B. L., & O'Brien, K. M. (1996). Course enrollment as self-regulatory behavior: Who takes optional high school math courses? *Learning and Individual Differences*, 8, 239–259. doi:10.1016/S1041-6080(96)90016-3

U.S. Census Bureau. (2006). *State and country quickfacts*. Retrieved from <http://quickfacts.census.gov/qfd/states/55000.html>

Vansteenkiste, M., Simons, J., Lens, W., Sheldon, K. M., & Deci, E. L. (2004). Motivating learning, performance, and persistence: The synergistic effects of intrinsic goal contents and autonomy-supportive contexts. *Journal of Personality and Social Psychology*, 87, 246–260. doi:10.1037/0022-3514.87.2.246

Walton, G. M., & Cohen, G. L. (2011, March 18). A brief social-belonging intervention improves academic and health outcomes of minority students. *Science*, 331, 1447–1451. doi:10.1126/science.1198364

Watt, H. M. (2005). Explaining gendered math enrollments for NSW Australian secondary school students. *New Directions for Child and Adolescent Development*, 2005, 15–29. doi:10.1002/cd.147

Watt, H. M., Eccles, J. S., & Durik, A. M. (2006). The leaky mathematics pipeline for girls: A motivational analysis of high school enrolments in Australia and the USA. *Equal Opportunities International*, 25, 642–659. doi:10.1108/02610150610719119

Watt, H. M. G., Shapka, J. D., Morris, Z. A., Durik, A. M., Keating, D. P., & Eccles, J. S. (2012). Gendered motivational processes affecting high school mathematics participation, educational aspirations, and career plans: A comparison of samples from Australia, Canada, and the United States. *Developmental Psychology*, 48, 1594–1611. doi:10.1037/a0027838

Wood, D., Kurtz-Costes, B., & Copping, K. E. (2011). Gender differences in motivational pathways to college for middle class African American youths. *Developmental Psychology*, 47, 961–968. doi:10.1037/a0023745

Yee, D. K., & Eccles, J. S. (1988). Parent perceptions and attributions for children's math achievement. *Sex Roles*, 19, 317–333. doi:10.1007/BF00289840

Received May 11, 2013

Revision received March 25, 2014

Accepted March 31, 2014 ■

Prekindergarten Children's Executive Functioning Skills and Achievement Gains: The Utility of Direct Assessments and Teacher Ratings

Mary Wagner Fuhs
University of Dayton

Dale Clark Farran and Kimberly Turner Nesbitt
Vanderbilt University

An accumulating body of evidence suggests that young children who exhibit greater executive functioning (EF) skills in early childhood also achieve more academically. The goal of the present study was to examine the unique contributions of direct assessments and teacher ratings of children's EF skills at the beginning of prekindergarten (pre-k) to gains in academic achievement over the pre-k year. Data for the current study come from a subsample of children recruited for a large-scale pre-k curriculum intervention. This subsample ($n = 719$) was restricted to all children who were native English speakers and had at least 1 pretest and posttest score on the assessments. Several important findings emerged. Teacher reports of EF and direct assessments were correlated, particularly when EF direct assessments were modeled as a single component score. When entered into the models simultaneously, both teacher ratings and direct assessments significantly predicted academic gains in literacy and mathematics; however, the direct assessments were only marginal in predicting gains in language. EF skills accounted for the largest proportion of variance in mathematics achievement gains. The value of using both types of measures in future research is discussed.

Keywords: executive function, prekindergarten, achievement, teacher ratings, direct assessments

An accumulating body of evidence suggests that young children who exhibit greater self-regulation abilities in early childhood achieve more academically (e.g., Blair & Razza, 2007; Bodovski & Farkas, 2007; Duncan et al., 2007; Li-Grining, Votruba-Drzal, Maldonado-Carreño, & Haas, 2010), have lower rates of hyperactive and disruptive behaviors (e.g., Espy, Sheffield, Wiebe, Clark, & Moehr, 2011; Séguin, Nagin, Assaad, & Tremblay, 2004), and are less likely to commit crimes and engage in delinquent behavior as adolescents or adults (Moffitt et al., 2011). Within the group of studies cited above are ones that capitalized on global ratings of children's self-regulation, including those asking parents and teachers to rate children's self-control, impulsivity, emotion regulation, persistence, and attention (e.g., Duncan et al., 2007; Moffitt et al., 2011). Other researchers, like Blair and Razza (2007),

have focused on direct child assessments to capture specific elements of children's self-regulation as they relate to school readiness and academic achievement.

In the current study, we focus on a set of skills within the domain of self-regulation that is typically referred to as executive functioning (EF) or cognitive control skills, including areas such as working memory, inhibitory control, and attention flexibility, and the contributions that both teacher reports and direct assessments of EF make to academic achievement. We examined the associations between children's EF skills and learning in prekindergarten (pre-k). Early childhood is a time when not only are children's EF skills showing rapid improvement (see Carlson, 2005; Garon, Bryson, & Smith, 2008), but also children are beginning to gain exposure to early academic concepts. Hence, this is a key developmental period in which to examine more closely the longitudinal associations between children's EF and early academic skills, providing findings that could inform assessment and intervention efforts in classrooms for young children.

Executive Functioning and Academic Achievement

EF skills include working memory (the ability to keep information active in memory and manipulate it), inhibitory control (the ability to inhibit salient but irrelevant information in favor of relevant information), and attention flexibility (the ability to shift and persist in attention; Garon et al., 2008). One or more of these EF skills have been positively associated with higher academic performance in young children in a variety of studies (e.g., Blair & Razza, 2007; Bodovski & Farkas, 2007; Duncan et al., 2007; Fuhs, Nesbitt, Farran, & Dong, 2014; Li-Grining et al., 2010). It has been hypothesized that stronger EF skills may allow children to meet the demands of an early childhood classroom better by facilitating

This article was published Online First July 28, 2014.

Mary Wagner Fuhs, Department of Psychology, University of Dayton; Dale Clark Farran and Kimberly Turner Nesbitt, Peabody Research Institute, Vanderbilt University.

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305E090009, awarded to Dale Clark Farran, Mark W. Lipsey, and Sandra J. Wilson. Mary Wagner Fuhs and Kimberly Turner Nesbitt were supported by an Institute of Education Postdoctoral Fellowship (R305B100016), awarded to Dale Clark Farran and Mark W. Lipsey. The opinions expressed are those of the authors and do not necessarily represent views of the Institute or the U.S. Department of Education. Special thanks to Deanna Meador and the PRI research team for their management of the data collection for this project.

Correspondence concerning this article should be addressed to Mary Wagner Fuhs, Department of Psychology, University of Dayton, 300 College Park, Dayton, OH 45469. E-mail: mfuhs1@udayton.edu

attention, memory for class rules, and engagement in academic content, all of which may allow them to benefit from an academic environment. EF skills may also work in a more direct way by aiding children's memory for salient information in early mathematics problem solving or increasing their flexibility in maintaining both letter sounds and symbols in memory during early literacy activities.

Much of the previous literature examining associations between EF skills and academic achievement in early childhood does not directly address the extent to which different methodologies for assessing EF skills may address the same construct and will relate to achievement growth. A growing literature focuses on using one or a battery of direct assessments to assess EF, whereas another line of research has addressed the associations of teacher reports of EF and academic skills development. A few studies have included both methodologies. In the following paragraphs, we address prior research on each methodology as well as when they have been used together before addressing gaps in the literature concerning the relative contribution each method makes to our understanding of the association between EF and growth in academic skills in young children.

Teacher Ratings of Executive Functioning Skills

Self-regulation typically serves as an umbrella term that includes cognitive and emotional components (Raver et al., 2012), associated with concepts in the teacher report literature such as "approaches to learning" and "self-control." Much of what we know about associations between self-regulation and academic achievement over time and in older children has derived from the use of these more global teacher report measures. For example, in the Early Childhood Longitudinal Study-Kindergarten, teachers rated children's approaches to learning, which included such behaviors as persistence at tasks, eagerness to learn, attention, learning independence, flexibility, and organization. Teacher ratings of these characteristics predicted mathematics achievement at all grade levels from kindergarten to second grade with the strongest effects for those children whose achievement fell in the bottom quartile (Bodovski, & Farkas, 2007).

Various assessments have been developed to capture self-regulation in young children through parent and/or teacher reports. For example, the Child Behavior Questionnaire (CBQ; Rothbart, Ahadi, Hershey, & Fisher, 2001) is considered a temperament scale. Based on Likert-type ratings of young children's emotional and cognitive regulation, the CBQ includes an inhibitory control subscale. This measure is typically defined as an assessment of effortful control, which includes both cognitive and emotion regulation components. Blair and Razza (2007) found that preschool teacher reports of effortful control using the CBQ were related to children's kindergarten mathematics and literacy skills. The coefficients were much weaker for the teacher reports when they were compared to direct child assessments of EF in preschool. The CBQ includes items reflecting both emotion regulation and cognitive regulation, making it difficult to compare directly the contributions of teacher reports of cognitive regulation alone and direct child assessments of targeted EF skills. In addition, Blair and Razza had no preschool measures of achievement; this intriguing study of the associations among these areas does not help us understand the

relationship between EF (assessed in different ways) and learning as measured by gains in academic skills across time.

While the CBQ includes both emotion and cognitive regulation, other teacher rating measures focus more specifically on EF skills. A clinically oriented measure, the Behavior Rating Inventory of Executive Function—Preschool Version (BRIEF-P; Gioia, Isquith, Retzlaff, & Espy, 2002), assesses whether children have deficits in particular areas of EF. The BRIEF-P has most commonly been used as a clinical and neuropsychological assessment of executive dysfunction. Other measures such as the Child Behavior Rating Scale (CBRS; Bronson, Tivnan, & Seppanen, 1995) focus more specifically on EF skills as they are manifested in typical classroom behavior in early childhood. Recent work has linked teacher ratings on the CBRS to children's early academic skills development above and beyond a direct child assessment of behavioral regulation (Wanless, McClelland, Acock, Chen, & Chen, 2011). Wanless et al. (2011) focused on the links between EF and achievement across different cultures. The timing of the teacher reports differed such that teachers in the United States did not complete teacher ratings until the middle of the school year. Thus, it is unclear the extent to which the differing timing of the teacher and direct assessments affected their contributions to academic achievement. The CBRS was also used in a kindergarten study that included a direct assessment of EF (Head, Toes, Knees, and Shoulders [HTKS]) as a predictor of achievement (Matthews, Ponitz, & Morrison, 2009); this sample was primarily middle-income and white. While achievement and HTKS were measured at pre- and posttest, teacher CBRS ratings were only collected in the spring and only for about 60% of the sample. Nevertheless, on this subset of children, both spring teacher ratings and fall HTKS scores were related to children's gains in math achievement over the year even when both were in the model.

A differently constructed measure of EF in the classroom is the Work-Related Skills subscale of the Cooper-Farran Behavioral Rating Scale (WRS; Cooper & Farran, 1988, 1991). The CBQ, the BRIEF-P, and the CBRS are similar to each other in that the items in each are rated on a Likert-type scale from, for example, "extremely untrue" to "extremely true" or "often" to "never." The WRS is different in that it assesses children's EF skills in academic learning contexts through the use of behaviorally anchored items related to classroom expectations. For example, one item lists "Listening to Teacher Giving Instructions to Group," and the anchors are "Attends to the teacher without reminders," "Occasionally inattentive; attention is easily regained by a cue from teacher," "Can maintain attending behavior with frequent reminders from the teacher," and "Seems to ignore the teacher; is very distracted and distracting." This type of scale, in contrast to a Likert rating scale, is situationally specific. The "person-situation" debate is a robust one in psychology; a trait approach is useful for predicting behaviors "averaged over many situations, occasions, and responses" (Epstein & O'Brien, 1985, p. 532). In a classroom, however, rating scales can suffer from what has been called a "reference group" problem (Heine, Lehman, Peng, & Greenholtz, 2002). While most clearly evident in cross-cultural work, Heine et al. (2002) argued that the reference group issue applies for any groups that might possibly possess different referents for their ratings, such as teachers. One solution Heine et al. proposed, although not without limitations, is to create items with concrete, objective, response options such as behavioral anchors.

Several studies of classrooms in the United States have demonstrated an association between teacher reports of young children's EF skills assessed by the WRS and their academic achievement (e.g., McClelland, Acock, & Morrison, 2006; McClelland, Morrison, & Holmes, 2000; Speece & Cooper, 1990), but many have focused on the relationship in kindergarten. For example, McClelland et al. (2006) found that children's EF skills, as assessed by the WRS subscale, predicted their academic achievement across domains in kindergarten and continued to predict mathematics and literacy achievement out to second grade after controlling for covariates, including prior achievement. McClelland et al. (2006) also found associations between ratings of EF skills and academic achievement out to sixth grade.

Although the work summarized above points to an association between teacher ratings of children's classroom-related EF and academic achievement, at least two specific questions remain. First, are these associations apparent prior to kindergarten? Children are increasingly likely to experience academic instruction in pre-k classrooms, especially those associated with public schools and a learning agenda. Because the previous research with a more classroom-specific scale has explored these links primarily in kindergarten, it is important to investigate whether teacher reports of children's EF in pre-k classrooms will also capture children's EF skills as they relate to their academic growth. Second, do these associations hold even when accounting for children's performance on direct assessments of EF? Concern is often raised about bias in teacher ratings primarily relating to teacher judgments of behavior and learning problems (e.g., Berg-Nielsen, Solheim, Bel-sky, & Wichstrom, 2012; Mullola et al., 2012). The assumption is that these ratings will be less valid and reliable than direct assessments of behavior. As direct child measures of EF have been developed and more information has accrued on their reliability and validity, an issue is whether they could substitute for teacher ratings and present equal, or perhaps better, indications of children's learning-related characteristics.

Direct Child Assessments of Executive Function

As previously mentioned, one increasingly common approach to measuring EF skills in young children is to assess them directly with a battery of tasks to tap working memory, inhibitory control, and attention flexibility. For the most part, these tasks were developed first in psychological laboratories. Researchers have found concurrent associations between directly assessed EF skills and children's academic achievement in both literacy and mathematics across different grade levels (e.g., Allan & Lonigan, 2011; Best, Miller, & Naglieri, 2011; Bull, Espy, Wiebe, Sheffield, & Nelson, 2011; Bull & Scerif, 2001; St. Clair-Thompson & Gathercole, 2006). Bull and Scerif (2001) found that several direct assessments of pre-k children's EF were concurrently related to their mathematics skills (see also Bull et al., 2011). In a cross-sectional study, Best et al. (2011) found contemporaneous associations between individually assessed EF measures and achievement at each grade level from age 5 through high school with the strongest, most consistent relationship being with mathematics achievement. Finally, longitudinal research suggests modest correlations between pre-k children's EF skills and their growth in academic skills as well (e.g., Bull, Espy, & Wiebe, 2008; Clark, Pritchard, & Woodward, 2010; Fuhs et al., 2014; McClelland et al., 2007).

Gaps in Extant Literature

While the literature indicates that both direct assessments and teacher ratings of children's EF skills are positively associated with children's academic achievement, several questions remain. First, is there a significant association between teacher reports of EF and direct child assessments in a pre-k sample? McClelland et al. (2007) examined correlations between HTKS as a direct assessment of EF and teacher ratings of children's social and behavioral regulation, but this work has not been extended to the WRS and a broader range of EF direct child assessments.

Second, what are the benefits and unique contributions of these two methods of assessment? A recent review of performance-based measures and ratings of EF found only modest correlations between them when each was used in the same study (Toplak, West, & Stanovich, 2013). The authors concluded that the two types of measures were actually tapping different cognitive levels in the respondent. Direct assessments they argued provide evidence of the individual's available processes, while ratings provide evidence of how those processes may or may not be used in an actual setting. Also, direct assessments provide an understanding of children's EF skills in a controlled or neutral context because they are administered to children individually usually in a quiet space. It is not clear what relationship performance in the controlled setting will have with children's EF skills in an ecological context like a classroom. On the other hand, while teacher ratings of children's behaviors can be particularly beneficial to understand children's EF skills in authentic classroom environments, ratings could be influenced by other aspects of children's abilities and skills besides EF.

Using both types of assessments together in a multimethod approach could yield a more complete understanding of children's EF skills, "providing important and nonredundant information about an individual's efficiency and success in achieving goals" (Toplak et al., 2013, p. 138). Moreover, an understanding of the relation between the two methods of assessing EF is informative for research because (a) utilizing direct child assessments is not always a feasible option for researchers, and (b) teacher reports may not be appropriate as the only means of assessing children's abilities (e.g., in curriculum interventions that focus specifically on developing self regulation).

Current Study

The goal of the present study is to examine the contribution of direct assessments and teacher ratings of children's EF skills at the beginning of pre-k to predict children's gains in academic achievement over the pre-k year. Four research questions were examined: (a) Is children's performance on direct assessments of EF positively correlated with teachers' ratings of their EF skills in the classroom context? (b) Are teacher ratings of children's EF skills at the beginning of pre-k associated with the gains children make in literacy, language, and mathematics across the pre-k year? (c) Is children's performance on direct assessments of EF at the beginning of pre-k associated with the gains they make in literacy, language, and mathematics across the pre-k year? (d) Are direct assessments and teacher ratings of EF, when examined together in a single model, significantly and uniquely associated with children's literacy, language, and mathematics gains?

Method

Participants

Data for the current study are a subsample from a larger sample of children ($N = 1,145$) recruited for a large-scale randomized control trial (RCT) to evaluate the effectiveness of the *Tools of the Mind* curriculum (Bodrova & Leong, 2007; Farran, Wilson, & Lipsey, 2013). All assessments were administered in English; therefore, we removed nonnative English speakers ($n = 380$) from the current sample to eliminate confounds due to limited English proficiency. To be consistent with the analytic sample of the RCT, we also removed children who did not have at least one pre- or posttest (primarily due to moving prior to the spring assessments). The analytic sample for the current study, therefore, consisted of 719 English-speaking pre-k students who had at least one complete pretest and posttest measure. There were 695 children ($M_{age} = 54$ months; $SD_{age} = 4$ months) with complete demographic, child assessment, and teacher report data for this study. Children with missing data points did not differ from the analytic sample on any demographic or pretest measure ($p > .05$), and because the cases with missing data constituted less than 5% of the analytic sample, we only used available data for each analysis rather than conducting multiple imputation. Girls were 46% of the sample, and children came from varied racial/ethnic backgrounds (36% Black, 52% White, 5% Hispanic, and 7% other).¹ Sixteen percent of the sample had an individualized education program (IEP), which was included in analyses as a covariate. Although precise socioeconomic status (SES) information was not available due to Family Educational Rights and Privacy Act regulations, all children in this study came from public pre-k programs targeted to low-income families. Therefore, it can be assumed that most, if not all, children in the study were from low-income backgrounds.

Children in the sample were nested in 80 classrooms in 57 schools in six school systems in the Southeastern United States. On average, nine children from each classroom were in the analysis sample. The average number of years of experience teaching pre-k for the teachers in the study was six. Because these data were drawn from an RCT, 32 classrooms were assigned to the *Tools of the Mind* condition and 28 classrooms were assigned to "business as usual," which involved a variety of curricula, but primarily *Creative Curriculum*, *Opening the World of Learning*, and *Building Blocks*. All analyses of main effects for curriculum on individual academic achievement measures and possible interactions between curriculum and demographic characteristics or children's pretests were nonsignificant in the RCT (Farran et al., 2013). Nonetheless, we included condition as a control variable in all of the present analyses as the experimental curriculum could potentially account for variance in academic achievement and EF in this sample.

Measures

Teacher reports. Children's classroom-specific EF skills as well as their more general social skills were assessed using the *Cooper-Farran Behavioral Rating Scale* (CFBRS; Cooper & Farran, 1988, 1991). The CFBRS is an assessment of young children's behaviors at school entry and consists of two subscales: Work-Related Skills (WRS) and Interpersonal Skills (IPS). The

WRS subscale rates children's EF skills as they are manifested in the classroom. The WRS consists of 16 items related to children's independent work, compliance and memory for instructions, and ability to complete tasks. The IPS subscale rates children's social skills. The IPS consists of 21 items related to children's ability to engage effectively in interactions with peers and teachers. All CFBRS items are rated from 1 to 7 using behavioral anchors distinctive to each odd-numbered item. As reported in the manual (Cooper & Farran, 1991), the test-retest reliability after an 8-week delay for the WRS and IPS subscales were .66 and .69, respectively. Interrater reliability on the measure has also been established between teacher and teacher aids; intraclass correlations between the raters were .79 and .78 for the WRS and IPS subscales. The two subscales also indicate high internal consistency (Chronbach's $\alpha > .94$).

Direct child assessments of EF. In the current study, EF was assessed using multiple measures that were chosen to cover the range of EF skills discussed in early childhood literature (see Garon et al., 2008). We included working memory, inhibitory control, and attention flexibility tasks (for additional information on each task, see <https://my.vanderbilt.edu/toolsofthemindvaluation/>), although each task naturally also tapped other abilities such as motor skills or language ability. This has commonly been called the "task impurity" problem as EF tasks not only tap non-EF skills but also typically tap more than one EF skill (Miyake, Friedman, Emerson, Witzki, & Howerter, 2000). To account for this, we were consistent with the methodology of prior research in this area and used a battery of EF tasks to create a composite score, drawing on the common EF variance shared by individual tasks (e.g., Wiebe, Espy, & Charak, 2008). Previous research with pre-k children appears to support a one-factor model of EF (e.g., Fuhs & Day, 2011; Hughes & Ensor, 2011; Wiebe et al., 2008; although see Miller, Giesbrecht, Muller, McInerney, & Kerns, 2012). We did, however, analyze the EF tasks both as individual tasks as well as a composite score to more fully capture their contributions to academic skills.

Visuo-spatial short-term memory and working memory were assessed using the *Corsi Blocks* task (Berch, Krikorian, & Huha, 1998; Corsi, 1972). We chose a visuospatial task instead of a verbal task because previous research has suggested that young children have great difficulty with verbal working memory tasks such as digit span (e.g., Bull et al., 2008). We also were concerned about the potential confounds of using a digit span task to predict mathematics skills because this task also taps digit familiarity. *Corsi Blocks* required children to point to a series of block patterns (block placement modeled after Berch et al., 1998) that became progressively more difficult by increasing the number of points in the pattern. Children were asked to repeat a pattern exactly as presented (Forward) and then to reverse a presented pattern (Backward). The experimenter tapped the blocks at a rate of approximately one block per second, and children received up to two attempts to successfully complete each span length until they scored incorrectly on both trials for a particular span length. Two

¹ Children's ethnicity was provided to us by schools from parent reports the schools collected. However, the reliability of this self-reported information was unclear, and thus, this variable was not used in analyses.

blocks of forward trials were conducted, followed by two blocks of backward trials.

Children received two practice trials prior to assessments in which the child only had to touch the same block as an experimenter to ensure that the child could perform the basic action required of the task. Then, the child received additional practice trials that were identical to test trials but that were followed by feedback. Following practice, the test trials without feedback began. Some have referred to the forward component of the task as a simple working memory task and the backward component as a complex working memory task (e.g., Garon et al., 2008), whereas others have referred to the forward component as a passive working memory task and the backward component as an active working memory task (e.g., Passolunghi & Cornoldi, 2008). Across interpretations, however, *Corsi Blocks* has been conceptualized as a working memory task and has been found to be significantly associated with academic skills in older children and adults, and particularly mathematics skills (see Raghubar, Barnes, & Hecht, 2010 for a review). The forward and backward versions of the task have shown high test-retest reliability in children ages 4 to 11 years ($r_s = .83$ and $.82$; Alloway, Gathercole, & Pickering, 2006).

Attention flexibility was measured with the *Dimensional Change Card Sort* (DCCS; Zelazo, 2006). The DCCS required children first to sort a set of cards according to one dimension (color) and then according to another (shape). Children were presented with two boxes, one with a red truck on it, and one with a blue star. Each box had a slit in the top for children to sort cards (e.g., blue trucks and red stars). The experimenter first demonstrated the color game on two trials and then conducted a rule check with the child ("Can you show me where the blue ones go in the color game?" and "Can you show me where the red ones go in the color game?"). Children received up to two trials for each rule check. Children were then instructed on each trial, "If it is a blue one, then put it here [pointing to blue star], but if it is a red one, put it here [pointing to red truck]." Children were given six trials and if they completed at least five of six correct, they moved on to the shape game. In the shape game, the rules were given without demonstration with cards, but children still had two opportunities to correctly complete the rule check. The same pass/fail criteria were used for both the color and shape sorts. If children completed both of these games successfully, children moved onto the advanced sort. In this game, children were asked to sort by color if they were presented with a card with a border around it and to sort by shape if the card had no border. Following both demonstration trials and rule checks, children completed 12 advanced sort trials, with nine out of 12 correct counted as passing.

Again, as with the other games, the rules were repeated on each trial. Using scoring suggested by Zelazo (2006), children received a score of 0 if they were unable to successfully sort five of six trials on the first dimension of color, a 1 if they sorted by the first dimension but did not meet the five-out-of-six cutoff criterion for sorting by the dimension of shape, a 2 if they were successful sorting by shape, and a 3 if they also passed the border sort (i.e., correct on at least nine of the 12 trials). Performance on DCCS has shown moderate test-retest reliability with children 36 to 72 months ($r = .44$; Müller, Kerns, & Konkin, 2012). Larger scores on the DCCS were interpreted as indicating greater attention flexibility.

Copy Design (Osborne, Butler, & Morris, 1984) required children to copy eight simple geometric shapes of increasing difficulty. Tasks of this type are drawing increasing attention (Cameron et al., 2012; Potter, Mashburn, & Grissmer, 2013). Cameron et al. (2012) described the task as requiring children "to process visual information from an external stimulus, invoke a mental representation, and coordinate motor movements to reproduce the image" (p. 1240). Children had two attempts to successfully draw each shape and each attempt was coded to indicate whether the child successfully replicated a design.

This task was used in the British Longitudinal Study analyzed by Duncan et al. (2007) and was one of the stronger long-term predictors of child outcomes. It was also recently used in a large-scale measurement study of children's cognitive self-regulation development (Lipsey et al., 2014). In this longitudinal measurement study, *Copy Design* was significantly correlated with a battery of cognitive self-regulation assessments and showed both construct validity via confirmatory factor analysis (Fuhs & Turner, 2012) and predictive validity for academic achievement (Lipsey et al., 2014). Interrater reliability for *Copy Design* in this study was established by two independent raters double coding 20% of the measures. The kappa coefficients for the 8 shapes ranged from .66 to 1.00 ($M_{kappa} = .79$). Each shape was scored 0 if coded as incorrect and 1 if coded as correct. Larger scores on the *Copy Design* were interpreted as indicating greater sustained attention.

Inhibitory control was assessed with *Peg Tapping* (PT; Diamond & Taylor, 1996) and *Head Toes Knees Shoulders* (HTKS; Ponitz, McClelland, Matthews, & Morrison, 2009). PT requires children to tap a peg once when an examiner taps it twice and to tap twice when an examiner taps once. Children completed 16 test trials that were scored 0 for incorrect responses and 1 for correct. If the child could not successfully complete practice trials on PT, he or she scored -1 and did not complete the test trials. Performance on PT has shown high test-retest reliability in 5-year-olds ($r = .74$; Nampijja et al., 2010).

HTKS is another task primarily assessing inhibitory control, although the task likely also taps children's working memory as directions are not repeated on each trial, attention shifting as the rules change during the game, and gross motor coordination. Children were asked to touch their heads when an examiner says "touch your toes" and to touch their toes when an examiner says "touch your head." If children were successful at inhibiting the prepotent response of behaving consistently with the prompt then two new prompts are added, children were then required to touch their knees when an examiner says "touch your shoulders" and vice versa. Children received up to a total of six practice trials and 20 test trials, and each trial was scored with 0 for an incorrect response, 1 for motion toward the incorrect response but ending with a correct response, and 2 for a correct response. Performance on HTKS has shown high test-retest reliability in 4-year-olds ($r = .80$; Meador, Turner, Lipsey, & Farran, 2013). Larger scores on both PT and HTKS indicated greater ability to inhibit a prepotent response.

Academic achievement. Academic achievement was assessed by administering seven subscales of the Woodcock Johnson III achievement battery (WJ-III; Woodcock, McGrew, Mather, 2001). Literacy skills were assessed with the *Letter-Word Identification* and *Spelling* subtests. *Letter-Word Identification* measures children's ability to identify and pronounce alphabet letters and

read words, while *Spelling* measures children's ability to draw simple shapes and write orally presented letters and words. Language skills were assessed with the *Academic Knowledge*, *Oral Comprehension*, and *Picture Vocabulary* subtests. *Academic Knowledge* tests children's factual knowledge of science, social studies, and the humanities; for young children the subtest mainly consists of labeling and identifying pictures, thus heavily relying on vocabulary. *Oral Comprehension* asks children to complete an orally presented passage by providing the appropriate missing word on the basis of semantic and syntactic cues. *Picture Vocabulary* asks children to name objects presented in pictures; it is a test of nouns or knowing the names of things. Mathematics skills were assessed with the *Applied Problems* and *Quantitative Concepts* subtests. *Applied Problems* measures children's ability to solve numerical and spatial problems accompanied by pictures while *Quantitative Concepts* measures children's understanding of number identification, sequencing, shapes, and symbols and in a separate section to manipulate the number line. All analyses were conducted using the item response theory-scaled W-Scores. Standard scores normed with a mean of 100 ($SD = 15$) can be more interpretable for descriptive purposes and are therefore presented in addition to the W score means in Table 1.

Procedure

Teacher reports were completed in the fall after children had acclimated to the classrooms, about 4–6 weeks. The ratings were collected close to the same time period in which children com-

pleted direct assessments. All direct assessments of children were conducted in a quiet area of the building in which they had their prekindergarten program. Data from children's EF assessments in the fall and their academic achievement in both fall and spring were used in analyses. Assessments were administered in a fixed order at each time point. In one child testing session, children completed PT, HTKS, and *Copy Design*, followed by the WJ-III *Oral Comprehension*, *Applied Problems*, *Quantitative Concepts*, and *Picture Vocabulary* subtests. The other testing session consisted of the DCCS and *Corsi Blocks*, followed by the WJ-III *Letter-Word Identification*, *Academic Knowledge*, and *Spelling* subtests. The average interval between fall and spring sessions was 7.38 months ($SD = 0.55$ months).

Analytic Approach

A series of multilevel models (children nested within classrooms, schools, and systems) was conducted to examine the associations between children's EF and academic achievement gains, using different methodologies to assess EF (teacher report and behavioral assessment). All predictors of interest and outcomes were included as standardized variables so that the parameter estimates could be compared across models. Prior to running conditional models, we first ran a fully unconditional model to determine the percentage of variance in academic achievement outcomes accounted for by the classroom, school, and system levels of our model. We then proceeded to run conditional models to test the associations between EF direct assessments and teacher

Table 1
Descriptive Statistics

Variable	N	M	SD	W score		Skew	t
				M	SD		
Corsi Forward T1	717	2.52	1.20			-0.61	
Corsi Backward T1	716	1.17	1.15			0.24	
DCCS T1	717	1.38	0.62			-0.14	
Copy Design T1	717	0.91	1.44			2.30	
HTKS T1	717	11.56	13.71			1.13	
Peg Tapping T1	717	5.31	5.86			0.37	
Work-Related Skills T1	716	4.60	1.13			-0.29	
Interpersonal Skills T1	716	5.18	1.06			-0.81	
Letter Word T1	716	93.23	12.21	318.66	24.57	0.09	
Letter Word T2	700	100.13	11.12	347.91	22.38	-0.23	37.32**
Spelling T1	716	79.15	12.55	336.40	23.52	-0.14	
Spelling T2	700	86.20	15.04	369.31	26.94	-0.26	36.24**
Academic Knowledge T1	716	91.31	12.61	436.21	15.82	-0.62	
Academic Knowledge T2	700	97.15	11.57	449.04	13.34	-0.66	29.20**
Oral Comprehension T1	717	94.55	11.15	445.39	13.42	-0.18	
Oral Comprehension T2	703	99.03	11.65	456.33	13.13	-0.15	27.13**
Picture Vocabulary T1	717	100.67	11.40	462.09	12.78	-2.30	
Picture Vocabulary T2	703	101.03	9.63	468.77	9.93	-2.47	16.93**
Applied Problems T1	717	96.69	12.53	390.27	25.10	-0.77	
Applied Problems T2	703	100.45	11.12	411.54	18.77	-0.88	29.90**
Quantitative Concepts T1	717	87.93	10.69	406.13	12.43	0.68	
Quantitative Concepts T2	703	92.97	12.87	422.50	14.60	-0.03	40.46**

Note. T1/T2 = Time 1/Time 2; DCCS = Dimensional Change Card Sort (Zelazo, 2006); HTKS = Head Toes Knees Shoulders (Ponitz, McClelland, Matthews, & Morrison, 2009). *t* statistics reported are from independent-samples *t* tests comparing assessment performance at T2 to T1. For Woodcock Johnson III achievement battery (WJ-III; Woodcock, McGrew, Mather, 2001), standard scores are reported, but W scores are also reported as they were used for tests of skewness and longitudinal analyses including *t* tests. Cooper-Farran Behavioral Rating Scale (CFBRS; Cooper & Farran, 1988, 1991). Work-Related Skills and Interpersonal Skills scores were based on average ratings on a 7-point scale.

** $p < .01$.

reports and gains in academic skills in literacy, language, and mathematics. Our predictors of interest were entered as fixed effects. A number of covariates were also entered as fixed effects at the child level including pretest scores, pre-post testing interval, age at pretest, gender, and IEP status. Because these data were taken from a large-scale RCT, condition was included as a fixed effect at the school level, although our primary interest was not in evaluating condition but rather in accounting for potential variance in outcomes due to variations in curricula used in classrooms. All multilevel models were run in IBM SPSS (Version 20 Mixed Models) using restricted maximum-likelihood estimation. A sample model equation for the EF direct assessments and teacher reports entered simultaneously to predict language outcomes is presented in the Appendix.

Results

Descriptive Statistics

Descriptive statistics for all variables are presented in Table 1. EF direct assessment scores and teacher ratings from the fall are presented in the top of the table; it is clear that for all the direct assessments variation among the children was great. Children entered pre-k with quite different EF skills. As indicated by the *t* tests in Table 1, children significantly improved their academic skills, with children making particularly large gains on W scores in WJ-III *Letter-Word Identification*, *Spelling*, and *Quantitative Concepts* subtests. We also examined each variable for evidence of nonnormality; both *Copy Design* and *Picture Vocabulary* showed evidence of skewness (see Table 1). *Copy Design* was positively skewed, and *Picture Vocabulary* was negatively skewed at both time points. Therefore, we performed log transformations of these variables before entering them in analyses. The log transformation reduced the skewness of *Copy Design* to .958, the skewness of *Picture Vocabulary* at Time 1 to -1.464 , and the skewness of *Picture Vocabulary* at Time 2 to -1.057 . Transformed variables were used in all subsequent analyses.

Data Reduction for EF and Achievement

Due to the nature and complexity of direct assessments of EF in young children, we are presenting results of our analytic models in two ways: (a) for the individual EF tasks entered simultaneously and (b) for the component EF score. We used principal component analysis (PCA) to extract common variance among the EF tasks (*Corsi Blocks*, *DCCS*, *Copy Design*, *HTKS*, and *Peg Tapping*) and saved a component score for one set of analyses. Using eigenvalues of >1 as the criterion to determine the number of components, a one-component PCA solution for EF at Time 1 accounted for 41.58% of the variance in the assessments. Component loadings for the EF measures were all above .50. We saved the component scores as a variable, which was standardized with a mean of 0 and a standard deviation of 1 and used in one set of analyses.

PCAs were also conducted to reduce redundancy in the measurement of academic achievement and to ensure that high correlations among the subtests would not jeopardize model specificity. We saved component scores for literacy (WJ-III *Letter-Word Identification* and *Spelling*), language (WJ-III *Academic Knowledge*, *Oral Comprehension*, and *Picture Vocabulary*), and mathe-

matics (WJ-III *Applied Problems* and *Quantitative Concepts*) at each time point. For literacy, we entered the WJ-III *Letter-Word Identification* and *Spelling* subtests into a PCA analysis, and we found a one-factor solution that accounted for 71.21% of the variance in the measures at Time 1. Again, for this and all other component scores, we saved these scores as variables and used them in analyses. At Time 2, the one-factor PCA solution for literacy with the same two measures accounted for 77.81% of the variance in the assessments. For language, the component score explained 72.24% of the variance at Time 1 and 72.53% of the variance at Time 2. The mathematics PCA produced a component that explained 82.16% of the variance in the assessments at Time 1 and 83.54% of the variance at Time 2. Within each PCA, the loadings of each of the measures onto the component were all above .80.

Correlations

Zero-order correlations among demographics, teacher ratings, the EF component score, individual EF direct assessments, and the achievement composites for both fall and spring are presented in Table 2. All academic achievement component scores, teacher ratings, and EF scores were moderately to strongly correlated with each other. Particularly strong correlations emerged between the EF direct assessments and the mathematics composite score at both time points. The EF direct assessment composite and teacher reports of EF (WRS) were strongly correlated, suggesting they are tapping a similar underlying construct. Interestingly the correlations between the EF composite score and entering achievement were somewhat higher than between teacher WRS ratings and entering academic skills, suggesting teachers were rating observed classroom behavior and not just children's skill levels. The correlations were notably weaker between teacher reports of social skills (IPS) and both direct assessments of EF and academic achievement.

Unconditional Multilevel Models

We first ran unconditional models to determine the percentage of variance in academic achievement gains in literacy, language, and mathematics that could be accounted for by the nesting levels of classroom, school, and system. It was especially important to establish the percentage of variance in academic achievement gains that could be accounted for by child-level differences as our predictors of interest were at the child level. If we found, for example, that the largest percentage of variance in achievement outcomes was at the classroom level, we would have little variance left to be explained by child-level predictors. Because our focus was on explaining children's pre-k gains in academic achievement, unconditional models included the covariates of gender, age, experimental condition, IEP status, interval of time that elapsed between pre- and posttest, and pretest achievement scores.

Based on the random parameters reported in the first columns of Tables 3, 4, and 5 (labeled "Unconditional Model"), 89.3% of variance in literacy outcomes, 94.1% of variance in language outcomes, and 92.7% of variance in mathematics outcomes (note that values were rounded) was attributed to child-level differences and could be modeled with our child-level predictors of interest, namely, EF direct assessments and teacher reports. Although the

Table 2
Correlations Among All Study Variables

Variable	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1. Corsi Forward	—														
2. Corsi Backward	.264**	—													
3. DCCS	.284**	.213***	—												
4. Copy Design	.280**	.199***	.176**	—											
5. HTKS	.334**	.235***	.249**	.249**	—										
6. Peg Tapping	.411**	.245***	.336***	.322**	.519***	—									
7. EF Factor	.680**	.518***	.586**	.554**	.717**	.775***	—								
8. WRS Rating	.384**	.257***	.291**	.271**	.350***	.393***	.511***	—							
9. IPS Rating	.196**	.092*	.137**	.074*	.175***	.195***	.232***	.649**	—						
10. Literacy T1	.413**	.231***	.289***	.485***	.351***	.445***	.575***	.413***	.146**	—					
11. Language T1	.393**	.284***	.422**	.250**	.491***	.498***	.617***	.458***	.188***	.492***	—				
12. Math T1	.447**	.324***	.398***	.378***	.523***	.591***	.701***	.493***	.200***	.630***	.713***	—			
13. Literacy T2	.358**	.187**	.236**	.396**	.235***	.349***	.458***	.431***	.171***	.667***	.433***	.542***	—		
14. Language T2	.355**	.275**	.375**	.198***	.417**	.440***	.546***	.462**	.191***	.429***	.810***	.632***	.447***	—	
15. Math T2	.471**	.275**	.381**	.348***	.444***	.528***	.645***	.486***	.190***	.612***	.634***	.755***	.653***	.653***	—
16. Gender	-.044	-.089*	-.111**	-.118**	-.080*	-.016	-.111**	-.227**	-.191**	-.144***	-.066	-.075*	-.214***	-.031*	-.051
17. Age	.220**	.108**	.060	.303**	.195**	.227**	.290**	.162**	-.013	.283***	.201**	.294***	.156**	.126**	.187***
18. Condition	.024	-.029	.029	.037	.013	.024	-.058	.065	.027	.023	.025	.023	-.035	-.013	
19. IEP Status	-.177**	-.070	-.112**	-.053	-.093*	-.150**	-.176**	-.222**	-.106**	-.117**	-.228**	-.170**	-.128**	-.200**	-.147**
20. Pre-Post Interval	.074	.061	.114**	.023	.067	.085*	.111**	.094*	.066	.045	.108**	.074	-.003	.166**	.130***

Note. DCCS = Dimensional Change Card Sort (Zelazo, 2006); HTKS = Head Toes Knees Shoulders (Ponitz, McClelland, Matthews, & Morrison, 2009); EF = executive functioning; WRS = Work-Related Skills subscale of the Cooper-Farran Behavioral Rating Scale (Cooper & Farran, 1988, 1991); IPS = Interpersonal Skills subscale of the Cooper-Farran Behavioral Rating Scale (Cooper & Farran, 1988, 1991); IEP = individualized education program; T1/T2 = Time 1/Time 2.

* $p < .05$. ** $p < .01$.

Table 3

EF Direct Assessments and Teacher Reports Predict End of Prekindergarten Literacy Skills

Variable	Unconditional model		Model 1		Model 2		Model 3		Model 4		Model 5	
	β	SE	β	SE	β	SE	β	SE	β	SE	β	SE
Fixed Parameters												
Intercept	0.01	0.08	0.00	0.09	0.00	0.09	-0.01	0.08	-0.01	0.09	-0.02	0.09
Gender	-0.12**	0.03	-0.12**	0.03	-0.12**	0.03	-0.09**	0.03	-0.09**	0.03	-0.09**	0.03
Age	-0.04	0.03	-0.08**	0.03	-0.07*	0.03	-0.06†	0.03	-0.08**	0.03	-0.07*	0.03
Curriculum Condition	-0.04	0.07	-0.05	0.07	-0.04	0.07	-0.01	0.07	-0.02	0.07	-0.01	0.07
IEP Status	-0.04	0.03	-0.02	0.03	-0.02	0.03	-0.01	0.03	0.00	0.03	-0.01	0.03
Pre-Post Interval	-0.02	0.03	-0.03	0.04	-0.03	0.04	-0.04	0.04	-0.04	0.04	-0.04	0.04
Pretest	0.66**	0.03	0.57**	0.03	0.59**	0.03	0.60**	0.03	0.54**	0.03	0.57**	0.03
Corsi Forward			0.09**	0.03					0.07*	0.03		
Corsi Backward			0.01	0.03					-0.01	0.03		
DCCS			0.04	0.03					0.03	0.03		
HTKS			-0.02	0.03					-0.03	0.03		
Peg Tapping			0.07†	0.04					0.05	0.04		
Copy Design			0.09**	0.03					0.09**	0.03		
EF Composite Score					0.16**	0.04					0.10**	0.04
EF Teacher Report							0.18**	0.03	0.15**	0.03	0.15**	0.03
Random Parameters												
Child	0.49**	0.03	0.47**	0.03	0.48**	0.03	0.47**	0.03	0.46**	0.03	0.46**	0.03
Classroom	0.03	0.02	0.02	0.03	0.03	0.04	0.03	0.04	0.02	0.04	0.03	0.04
School	0.00	0.00	0.01	0.04	0.00	0.04	0.01	0.05	0.01	0.04	0.01	0.05
System	0.02	0.02	0.03	0.03	0.03	0.03	0.02	0.02	0.02	0.03	0.03	0.03
Pseudo- R^2			0.04		0.03		0.05		0.06		0.06	

Note. EF = executive functioning; IEP = individualized education plan; DCCS = Dimensional Change Card Sort (Zelazo, 2006); HTKS = Head Toes Knees Shoulders (Ponitz, McClelland, Matthews, & Morrison, 2009). Pseudo- R^2 estimates indicate the amount of within-child variability in end of prekindergarten literacy skills explained by the addition of EF measures to the unconditional model.

† $p < .10$. * $p < .05$. ** $p < .01$.

percentage of variance accounted for by the different levels of the model varied across academic content areas we decided to account for all levels in all academic content area analytic models to maintain consistency across models and to aid in comparison of effects.

Conditional Multilevel Models

In Tables 3–5, we present the associations between children's fall EF scores and their spring academic achievement in the areas of literacy, language and mathematics after controlling for demographic covariates and children's academic achievement in the fall, in other words the gain in achievement related to initial EF scores. Each table focuses on a different academic area and each contains five models. The first model in each table is the model with the EF direct assessments entered individually into the model to predict gains in academic achievement. The second model (Model 2), shows results when the EF composite score was entered alone. Model 3 in each table depicts the results when the WRS ratings are entered; Model 4 shows the individual direct assessment scores with the addition of the teacher ratings. Finally in each table, Model 5 examines the joint contribution of the EF direct assessment composite and the teacher ratings of EF in predicting gains in achievement across the year.

Multilevel models for EF direct assessments. First we discuss the results from Models 1 and 2 examining the effects for the individual direct assessments entered simultaneously versus the composite score alone. For literacy gains (Table 3), *Corsi Forward*, *Peg Tapping*, and *Copy Design* were significant or marginal

predictors. For language gains shown in Table 4, none of the individual EF measures significantly predicted growth except for a marginal effect for *Corsi Backward*. The individual significant EF predictors of mathematics gains as shown in Table 5 were *Corsi Forward*, *Peg Tapping*, *Copy Design* and *DCCS*. HTKS did not predict any achievement content area gains. When examining variance accounted for by adding the group of EF assessments to the conditional model, the pseudo- R^2 estimate of effect size was largest for the mathematics content area.

Model 2 in Tables 3–5 presents the results when EF skills are entered as a PCA composite score. The composite EF score was predictive of children's literacy, language, and mathematics skills in the spring with fall pretest scores entered as a covariate. Again, the pseudo- R^2 estimate of effect size was the largest for mathematics. Thus, Models 1 and 2 indicate that direct assessments of EF both individually and as a composite related to achievement gains over and above children's entering skill levels, but the magnitude of effects varied both by individual EF assessment and by academic content area.

Multilevel models for EF teacher reports. Model 3 tested the association between teachers' ratings of children's EF and spring academic achievement in the areas of literacy, language, and mathematics after controlling for demographic covariates and children's academic achievement in the fall. After accounting for covariates and academic pretest skills, Model 3 in each of Tables 3–5 demonstrates that teachers' fall reports of EF significantly predicted literacy, language, and mathematics outcomes, with the pseudo- R^2 estimate of effects size for literacy achievement being

Table 4

EF Direct Assessments and Teacher Reports Predict End of Prekindergarten Language Skills

Variable	Unconditional model		Model 1		Model 2		Model 3		Model 4		Model 5	
	β	SE	β	SE	β	SE	β	SE	β	SE	β	SE
Fixed Parameters												
Intercept	0.07	0.04	0.07	0.04	0.07 [†]	0.04	0.06	0.04	0.06	0.04	0.06	0.04
Gender	0.02	0.02	0.03	0.02	0.03	0.02	0.05*	0.02	0.05*	0.02	0.05*	0.02
Age	-0.03	0.02	-0.04	0.02	-0.05	0.03	-0.04 [†]	0.02	-0.04 [†]	0.02	-0.05*	0.02
Curriculum Condition	-0.11*	0.05	-0.11*	0.05	-0.12	0.06	-0.09 [†]	0.05	-0.10 [†]	0.05	-0.10 [†]	0.05
IEP Status	-0.03	0.02	-0.02	0.02	-0.02	0.02	-0.01	0.02	-0.01	0.02	-0.01	0.02
Pre-Post Interval	0.07**	0.03	0.06*	0.03	0.06*	0.03	0.06*	0.02	0.06*	0.03	0.06*	0.03
Pretest	0.80**	0.02	0.75**	0.03	0.75**	0.03	0.75**	0.03	0.73**	0.03	0.73**	0.03
Corsi Forward		0.03	0.03						0.01	0.03		
Corsi Backward		0.04 [†]	0.02						0.03	0.02		
DCCS		0.03	0.02						0.03	0.02		
HTKS		0.01	0.03						0.01	0.03		
Peg Tapping		0.03	0.03						0.02	0.03		
Copy Design		-0.01	0.02						-0.01	0.02		
EF Composite Score					0.09**	0.03					.05 [†]	0.03
EF Teacher Report							0.10**	0.04	0.10**	0.03	.10**	0.03
Random Parameters												
Child	0.32**	0.02	0.32**	0.02	0.32**	0.02	0.31**	0.02	0.31**	0.02	0.32**	0.02
Classroom	0.02 [†]	0.01	0.02 [†]	0.01	0.02 [†]	0.01	0.01	0.01	0.01 [†]	0.01	0.01 [†]	0.01
School	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
System	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Psuedo- R^2			0.01		0.01		0.02		0.02		0.02	

Note. EF = executive functioning; IEP = individualized education plan; DCCS = Dimensional Change Card Sort (Zelazo, 2006); HTKS = Head Toes Knees Shoulders (Ponitz, McClelland, Matthews, & Morrison, 2009). Psuedo- R^2 estimates indicate the amount of within-child variability in end of prekindergarten language skills explained by the addition of EF measures to the unconditional model.

[†] $p < .10$. * $p < .05$. ** $p < .01$.

the largest. Therefore, although EF direct assessments accounted for the most variance in mathematics achievement gains, EF teacher reports accounted for the most variance in literacy achievement gains.

Multilevel models for EF direct assessments and teacher reports together. Finally, we assessed the unique contributions of direct assessments of EF, individually and as a composite, and teacher reports of EF when entered into a model simultaneously (see Tables 3–5, Models 4 and 5). We again ran two separate models, one in which EF direct assessments were entered individually with the teacher ratings and one in which the EF direct assessments composite score was entered with teacher ratings but without the individual EF assessments. After controlling for covariates and academic pretests, teacher reports remained significant predictors of academic achievement when entered simultaneously with EF direct child assessments.

As shown in Table 3, Model 4, for literacy gains both teacher ratings and EF assessments of *Corsi Forward* and *Copy Design* continued to be significant predictors of literacy outcomes. Similarly both teacher ratings and the EF component score were uniquely related to literacy gains (Model 5). For language outcomes presented in Table 4, none of the individual EF measures was significantly associated with outcomes, but Model 5 shows that the EF component score was a marginal predictor when included in the model with teacher reports. Table 5 shows that *Corsi Forward*, *DCCS*, *Peg Tapping*, *Copy Design* (Model 4), and the EF component score (Model 5) were significant or marginal predictors of mathematics outcomes along with teacher ratings.

To summarize, Models 4 and 5 indicate that both EF methods (teacher reports and direct assessments, whether entered individually or as a composite) were significant predictors of children's achievement outcomes in language, literacy, and mathematics. The standardized coefficients for teacher reports were larger for language and literacy, whereas the coefficients for direct assessments were larger for mathematics. The magnitude of the EF effect when both assessment types were entered simultaneously was largest for mathematics outcomes.

Multilevel models for interpersonal skills. Despite the fact that our analyses related to *gains* in achievement by including pretest in the models, one could question whether teacher ratings reflected a general favorable bias toward higher achieving children. If so, that bias should have been reflected in all the ratings teachers provided, both the ones related to EF and the ones related to social interactions. Therefore, we also assessed whether teachers' fall ratings of their children from the other CFBRS scale, the IPS subscale, also accounted for unique variation in academic achievement gains across the pre-k year above and beyond variance accounted for by children's EF direct assessment as well as covariates including the academic skills pretests. Teacher ratings of children's interpersonal skills were not predictive of their academic achievement gains above and beyond the EF direct assessments when entered individually for literacy ($\beta = .04$, $SE = .03$, $p = .171$), language ($\beta = .03$, $SE = .02$, $p = .226$), or mathematics ($\beta = .003$, $SE = .02$, $p = .901$). Teacher ratings of children's interpersonal skills were also nonsignificant predictors of academic achievement gains when entered into models with the EF

Table 5

EF Direct Assessments and Teacher Reports Predict End of Prekindergarten Mathematics Skills

Variable	Unconditional model		Model 1		Model 2		Model 3		Model 4		Model 5	
	β	SE	β	SE	β	SE	β	SE	β	SE	β	SE
Fixed Parameters												
Intercept	0.05	0.06	0.05	0.05	0.05	0.05	0.04	0.06	0.04	0.05	0.04	0.05
Gender	0.00	0.02	0.02	0.02	0.02	0.02	0.03	0.02	0.04	0.02	0.04	0.02
Age	-0.02	0.03	-0.06*	0.03	-0.05*	0.03	-0.03	0.03	-0.06*	0.03	-0.06*	0.03
Curriculum Condition	-0.06	0.06	-0.08	0.06	-0.07	0.06	-0.04	0.06	-0.06	0.06	-0.06	0.06
IEP Status	-0.02	0.03	0.00	0.02	-0.01	0.02	0.00	0.02	-0.01	0.02	0.01	0.02
Pre-Post Interval	0.05	0.03	0.03	0.03	0.04	0.03	0.04	0.03	0.03	0.03	0.03	0.03
Pretest	0.74**	0.03	0.61**	0.03	0.60**	0.03	0.67**	0.06	0.60**	0.03	0.57**	0.03
Corsi Forward			0.12**	0.03					0.11**	0.03		
Corsi Backward			0.00	0.03					-0.01	0.02		
DCCS			0.07*	0.03					0.06*	0.03		
HTKS			0.02	0.03					0.02	0.03		
Peg Tapping			0.06†	0.03					0.05†	0.03		
Copy Design			0.07*	0.03					0.06*	0.03		
EF Composite Score					0.23**	0.03					0.20**	0.03
EF Teacher Report							0.14**	0.03	0.10**	0.03	0.10**	0.03
Random Parameters												
Child	0.39**	0.02	0.36**	0.02	0.36**	0.02	0.37**	0.02	0.36**	0.02	0.36**	0.02
Classroom	0.03*	0.01	0.02	0.02	0.03*	0.01	0.03*	0.01	0.02	0.02	0.03*	0.01
School	0.00	0.02	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.00
System	0.00	0.01	0.00	0.01	0.00	0.01	0.00	0.01	0.00	0.01	0.00	0.01
Pseudo- R^2			0.06		0.05		0.03		0.07		0.07	

Note. EF = executive functioning; IEP = individualized education plan; DCCS = Dimensional Change Card Sort (Zelazo, 2006); HTKS = Head Toes Knees Shoulders (Ponitz, McClelland, Matthews, & Morrison, 2009). Pseudo- R^2 estimates indicate the amount of within-child variability in end of prekindergarten mathematics skills explained by the addition of EF measures to the unconditional model.

* $p < .10$. * $p < .05$. ** $p < .01$.

composite score—literacy ($\beta = .04$, $SE = .03$, $p = .193$), language ($\beta = .03$, $SE = .02$, $p = .221$), mathematics ($\beta = .01$, $SE = .03$, $p = .829$).

Discussion

In the present study, we compared the unique contributions of teacher reports of EF and direct child assessments of EF when added into a model simultaneously to predict gains in academic skills across the pre-k year. Three important findings emerged. First, children's EF skills both as rated by teachers and as observed in direct child assessments were significantly related to each other. Second, when entered in separate models, direct assessments and teacher reports of children's EF skills at school-entry were significantly related to their academic gains in literacy, language, and mathematics in pre-k above and beyond covariates. Third, both teacher reports of children's EF and direct assessments of EF remained significant predictors of literacy and mathematic gains even when both were entered into a model simultaneously. Direct assessments of EF were only marginally associated with language gains when entered into the model simultaneously with teacher reports.

Teacher-reported EF was strongly related to behavioral assessments of children's EF, which suggests that they may be tapping similar if not identical underlying characteristics; our correlations were stronger than those summarized by Toplak et al. (2013) in clinical populations. In fact, as can be seen in the simple correlations, in some cases the teacher reports of EF were more highly correlated with individual EF direct assessments than some of the

correlations among EF direct child assessments themselves. This provides further evidence that these two different methodologies may be tapping into similar skill sets in young children and also speaks to the measurement issues with individual direct assessments. Although previous research has estimated correlations between the HTKS and teacher reports of behavioral regulation in the early childhood classroom (McClelland et al., 2007), the current study extends this research by creating a component score of EF that draws upon the common variance among these measures, and associating it with teacher reports of specific EF behaviors in the classroom. The correlations obtained in the current study were much higher than those obtained by McClelland et al. (2007) using only the HTKS, suggesting the possibility that when the variance unique to each individual assessment is removed, the overlap between teacher reports and child direct assessments of EF may be higher than previously reported. However, the correlations between teacher reports of EF and direct child assessments were not so high as to suggest complete redundancy in measurement. It is quite possible that the teacher report may tap skills that are not being tapped by direct child assessments, lending support to the idea that both methodologies may yield important information about young children's EF skills. Direct assessments may capture children's available cognitive processes and teacher reports may assess how these processes are used in a real-world setting.

We found positive associations between children's entering EF skills, assessed through direct child assessments and teacher reports, and their gains in literacy, language, and mathematics skills. When entered into models separately, the effect size estimates of

the unique variance accounted for by EF (above and beyond covariates and pretests) were larger for the model of teacher reports predicting literacy and language compared to the models including direct assessments. This was not the case for mathematics achievement gains, as the addition of direct assessments of EF to the model accounted for a larger proportion of variance compared to the addition of teacher reports. When examining teacher reports and direct assessments entered together as predictors, effects differed such that after controlling for covariates and pretests, the strongest effect of children's school-entry EF skills was observed for gains in mathematics achievement. Effect sizes were smaller across the board for the variance accounted for by EF above and beyond language pretests and covariates.

While the pseudo- R^2 estimate of effects size indicated that EF skills accounted for 7% of the variance in gains in children's mathematics achievement, it is necessary to interpret the magnitude of this effect based on its practical value for a given field (Cumming, 2014). In the present case, the practical significance of the effects must be interpreted in light of explaining unique variance in *gains* in children's academic achievement *beyond* that which can be explained by initial achievement as well as covariates. Such a conservative approach suggests that even for smaller pseudo- R^2 estimates, the effects remain meaningful for practice.

Several researchers have previously posited that mathematics skills uniquely tax children's inhibitory control, attentional flexibility, and working memory (Blair, Knipe, & Gamson, 2008). Specifically, recent accounts suggest that although children may initially heavily recruit EF resources for academic learning across domains, certain skills, such as literacy, may become more automatic requiring less higher order problem solving compared to mathematics tasks, which likely only increase in their cognitive demands as new mathematical concepts are learned (e.g., Blair et al., 2008; Welsh et al., 2010). Our results are consistent with this account when examining the variance in mathematics achievement gains accounted for by both EF direct child assessments and teacher report simultaneously, such that the effect size values were the largest for mathematics models. However, we found somewhat smaller but still important predictions from EF measures to gains in literacy, suggesting that during the pre-k year at least, these skills are not as automatic as they will become in kindergarten and first grade.

When examining direct child assessments and teacher reports separately we found that teacher reports of EF accounted for more variance in literacy and language achievement gains than did EF direct assessments examined alone. Conversely, the model with EF direct assessments alone accounted for more variance in mathematics achievement gains compared to the model with EF teacher reports alone. Prior work on EF and achievement is largely based on the use of direct child assessments; the current study results suggest that the addition of teacher reports will yield a more comprehensive picture, at least in pre-k. It could be that previous findings were limited by the use of measures of only one methodology, and perhaps that literacy, language, and mathematics skills may all be influenced by EF skills in the pre-k year but in different ways. For example, it could be the case that literacy and language skills are more affected by the types of skills tapped by teacher reports, namely, how children use their EF skills in classroom learning, whereas mathematics skills are more directly affected by the cognitive processes themselves. Future research is

necessary to examine these effects beyond pre-k and into early elementary school to determine if differential patterns emerge.

The difference in magnitude of effects of EF skills on literacy and mathematics gains compared to language gains is not necessarily surprising considering the nature of pre-k instruction in which literacy and mathematics skills receive more explicit attention compared to language skills. Thus, the benefits of having greater EF skills that allow children to attend and remain engaged during classroom instruction may be more relevant for early literacy and mathematics skills compared to language skills. Also, in this particular data set, children made substantially less gain in the language across the pre-k year, and the correlations from pre- to posttest were very high, suggesting strong stability and less intraindividual variability to be explained by EF measures.

It is worth commenting on the fact that we did not find HTKS to be a unique predictor of gains in any of the three academic areas, which is in contrast to several recent studies with this measure (e.g., McClelland et al., 2007; Wanless et al., 2011). A big difference in our study compared to the ones cited above is that we used several direct assessment measures and not just HTKS alone. It is apparent from Table 2 and the zero-order correlations that Peg Tapping and HTKS were the most highly correlated among the direct assessment measures ($r = .52$). Peg Tapping generally correlated more highly with the other direct assessments than HTKS did. It is possible that alone HTKS is an important contributor to achievement gains, but its contribution was swamped by the stronger relations between gains and some of the other EF direct assessment measures.

Taken together, these findings illustrate the potential for ecologically valid teacher ratings of EF skills in combination with direct assessments, especially with regard to examining the interrelations between EF skills and academic achievement in field-based research. Importantly, the same pattern of associations was not observed for teacher reports of children's social skills, suggesting that teachers were not simply more positively predisposed to some children than others and also as Duncan et al. (2007) also found, that social skills may be important but perhaps not for academic achievement. Teachers seem to be able to recognize specific types of behaviors in young children that will facilitate their learning in the classroom over the year. Researchers will sometimes use a battery of direct assessments when measuring the various aspects of children's EF but seldom does the battery include teacher reports. Such batteries can require taking children out of their classrooms for one-on-one testing for up to 45 min, which is not always desirable or feasible. Moreover, because traditional direct assessments of EF were initially developed for use in neurological research, many assessments of EF are not situated within the context of typical preschool learning activities. Teacher reports that are easily administered and ecologically valid could considerably enhance our understanding of the interrelations between the development of EF and early academic success, particularly when used in conjunction with direct child assessments.

Limitations

The contributions of this research notwithstanding, study limitations must be acknowledged. While the current findings

help us better understand the unique value of various modes of assessing young children's EF skills, the correlational design of the study does not permit causal conclusions regarding the associations between children's EF skills and gains in achievement over the pre-k year. In particular, while we demonstrated that teacher ratings were convergent with traditional direct assessments of EF and explained unique variance in children's literacy, language, and mathematics achievement in conjunction with direct assessments of EF, it is possible that teacher ratings could be capturing other aspects of children's scholastic abilities. It may be that future research could be conducted to further validate the use of teacher ratings by including statistical controls for other aspects of children's scholastic abilities that might be confounded with ratings of EF skills, including general intelligence. Another limitation with the use of the Work-Related Skills subscale of the Cooper-Farran Behavioral Rating Scale is that it does have three items that reference emotion or social context. Thus, it is possible that although the measure primarily assesses EF, it may also capture other additional skills as well. Future work should examine associations between the Work-Related Skills subscale and other assessments of emotion regulation to get a clearer picture of the extent to which the Work-Related Skills subscale might also capture more affectively laden components of self-regulation skills.

Conclusions

Both EF direct assessments and EF teacher reports explained unique variance in children's academic achievement gains in literacy and mathematics. Teacher reports may be an ecologically valid and efficient way of capturing children's EF development in early childhood classrooms when direct assessments are not feasible. If possible, including both teacher ratings and one or more direct assessments would seem to be the best course.

References

Allan, N. P., & Lonigan, C. J. (2011). Examining the dimensionality of effortful control in preschool children and its relation to academic and socioemotional indicators. *Developmental Psychology, 47*, 905–915. doi:10.1037/a0023748

Alloway, T. P., Gathercole, S. E., & Pickering, S. J. (2006). Verbal and visuospatial short-term and working memory in children: Are they separable? *Child Development, 77*, 1698–1716.

Berch, D. B., Krikorian, R., & Huha, E. M. (1998). The Corsi block-tapping task: Methodological and theoretical considerations. *Brain and Cognition, 38*, 317–338. doi:10.1006/brcg.1998.1039

Berg-Nielsen, T., Solheim, E., Belsky, J., & Wichstrom, L. (2012). Preschoolers' psychosocial problems: In the eyes of the beholder? Adding teacher characteristics as determinants of discrepant parent-teacher reports. *Child Psychiatry and Human Development, 43*, 393–413. doi:10.1007/s10578-011-0271-0

Best, J. R., Miller, P. H., & Naglieri, J. (2011). Relations between executive function and academic achievement from ages 5 to 17 in a large, representative national sample. *Learning and Individual Differences, 21*, 327–336. doi:10.1016/j.lindif.2011.01.007

Blair, C., Knipe, H., & Gamson, D. (2008). Is there a role for executive functions in the development of mathematics ability? *Mind, Brain, and Education, 2*, 80–89. doi:10.1111/j.1751-228X.2008.00036.x

Blair, C., & Razza, R. P. (2007). Relating effortful control, executive function, and false belief understanding to emerging math and literacy ability in kindergarten. *Child Development, 78*, 647–663. doi:10.1111/j.1467-8624.2007.01019.x

Bodovski, K., & Farkas, G. (2007). Mathematics growth in early elementary school: The roles of beginning knowledge, student engagement, and instruction. *The Elementary School Journal, 108*, 115–130. doi:10.1086/525550

Bodrova, E., & Leong, D. J. (2007). *Tools of the mind: The Vygotskian approach to early childhood education* (2nd ed.). Columbus, OH: Merrill/Prentice Hall.

Bronson, M. B., Tivnan, T., & Seppanen, P. S. (1995). Relations between teacher and classroom activity variables and the classroom behaviors of prekindergarten children in Chapter 1 funded programs. *Journal of Applied Developmental Psychology, 16*, 253–282. doi:10.1016/0193-3973(95)90035-7

Bull, R., Espy, K. A., & Wiebe, S. A. (2008). Short-term memory, working memory, and executive functioning in preschoolers: Longitudinal predictors of math achievement at age 7 years. *Developmental Neuropsychology, 33*, 205–228. doi:10.1080/87565640801982312

Bull, R., Espy, K. A., Wiebe, S. A., Sheffield, T. D., & Nelson, J. M. (2011). Using confirmatory factor analysis to understand executive control in preschool children: Sources of variation in emergent math achievement. *Developmental Science, 14*, 679–692. doi:10.1111/j.1467-7687.2010.01012.x

Bull, R., & Scerif, G. (2001). Executive functioning as a predictor of children's math ability: Inhibition, switching, and working memory. *Developmental Neuropsychology, 19*, 273–293. doi:10.1207/S15326942DN1903_3

Cameron, C., Brock, L., Murrah, W., Bell, L., Worzalla, S., Grissmer, D., & Morrison, F. (2012). Fine motor skills and executive function both contribute to kindergarten achievement. *Child Development, 83*, 1229–1244. doi:10.1111/j.1467-8624.2012.01768.x

Carlson, S. M. (2005). Developmentally sensitive measures of executive function in preschool children. *Developmental Neuropsychology, 28*, 595–616. doi:10.1207/s15326942dn2802_3

Clark, C. A. C., Pritchard, V. E., & Woodward, L. J. (2010). Preschool executive functioning abilities predict early mathematics achievement. *Developmental Psychology, 46*, 1176–1191. doi:10.1037/a0019672

Cooper, D. H., & Farran, D. C. (1988). Behavioral risk in kindergarten. *Early Childhood Research Quarterly, 3*, 1–19. doi:10.1016/0885-2006(88)90026-9

Cooper, D. H., & Farran, D. C. (1991). *The Cooper-Farran Behavioral Rating Scales*. Brandon, VT: Clinical Psychology.

Corsi, P. M. (1972). *Human memory and the medial temporal region of the brain* (Doctoral dissertation, McGill University). Retrieved from http://digitool.Library.McGill.CA:80/R/-?func=dbin-jump-full&object_id=93903&silo_library=GEN01

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science, 25*, 7–29. doi:10.1177/0956797613504966

Diamond, A., & Taylor, C. (1996). Development of an aspect of executive control: Development of the ability to remember what I said and to "do as I say, not as I do." *Developmental Psychobiology, 29*, 315–334.

Duncan, G. J., Dowsett, C. J., Claessens, A., Magnuson, K., Huston, A. C., Klebanov, P., . . . Brooks-Gunn, J. (2007). School readiness and later achievement. *Developmental Psychology, 43*, 1428–1446. doi:10.1037/0012-1649.43.6.1428

Epstein, S., & O'Brien, E. (1985). The person-situation debate in historical and current perspective. *Psychological Bulletin, 98*, 513–537. doi:10.1037/0033-295X.98.3.513

Espy, K. A., Sheffield, T. D., Wiebe, S. A., Clark, C. A. C., & Moehr, M. J. (2011). Executive control and dimensions of problem behaviors in preschool children. *Journal of Child Psychology and Psychiatry, 52*, 33–46. doi:10.1111/j.1469-7610.2010.02265.x

Farran, D. C., Wilson, S. J., & Lipsey, M. W. (2013, March). *Effects of a curricular attempt to improve self-regulation and achievement in pre-kindergarten children*. Paper presented at the 2013 Society for Research in Child Development Conference, Seattle, WA.

Fuhs, M. W., & Day, J. D. (2011). Verbal ability and executive functioning development in preschoolers at head start. *Developmental Psychology, 47*, 404–416. doi:10.1037/a0021065

Fuhs, M. W., Nesbitt, K. T., Farran, D. C., & Dong, N. (2014). Longitudinal associations between executive functioning and academic achievement across content areas. *Developmental Psychology*. Advance online publication. doi:10.1037/a0036633

Fuhs, M. W., & Turner, K. A. (2012, February). *Evaluating group and longitudinal measurement equivalence in a battery of cognitive self-regulation measures for preschoolers*. Poster presented at the Society for Research in Child Development 2012 themed meeting: Developmental methodology, Tampa, FL.

Garon, N., Bryson, S. E., & Smith, I. M. (2008). Executive function in preschoolers: A review using an integrative framework. *Psychological Bulletin, 134*, 31–60. doi:10.1037/0033-2909.134.1.31

Gioia, G. A., Isquith, P. K., Retzlaff, P. D., & Espy, K. A. (2002). Confirmatory factor analysis of the Behavior Rating Inventory of Executive Function (BRIEF) in a clinical sample. *Child Neuropsychology, 8*, 249–257. doi:10.1076/chin.8.4.249.13513

Heine, S., Lehman, D., Peng, K., & Greenholtz, J. (2002). What's wrong with cross-cultural comparisons of subjective Likert scales? The reference-group effect. *Journal of Personality and Social Psychology, 82*, 903–918. doi:10.1037/0022-3514.82.6.903

Hughes, C., & Ensor, R. (2011). Individual differences in growth in executive function across the transition to school predict externalizing and internalizing behaviors and self-perceived academic success at 6 years of age. *Journal of Experimental Child Psychology, 108*, 663–676. doi:10.1016/j.jecp.2010.06.005

Li-Grining, C. P., Votruba-Drzal, E., Maldonado-Carreño, C., & Haas, K. (2010). Children's early approaches to learning and academic trajectories through fifth grade. *Developmental Psychology, 46*, 1062–1077. doi:10.1037/a0020066

Lipsey, M. W., Nesbitt, K. T., Farran, D. C., Dong, N., Fuhs, M. W., & Wilson, S. J. (2014). *Learning-related cognitive self-regulation school readiness measures for preschool children: Optimizing predictive validity for achievement*. Manuscript submitted for publication.

Matthews, J. S., Ponitz, C., & Morrison, F. (2009). Early gender differences in self-regulation and academic achievement. *Journal of Educational Psychology, 101*, 689–704. doi:10.1037/a0014240

McClelland, M. M., Accock, A. C., & Morrison, F. J. (2006). The impact of kindergarten learning-related skills on academic trajectories at the end of elementary school. *Early Childhood Research Quarterly, 21*, 471–490. doi:10.1016/j.ecresq.2006.09.003

McClelland, M. M., Cameron, C. E., Connor, C. M., Farris, C. L., Jewkes, A. M., & Morrison, F. J. (2007). Links between behavioral regulation and preschoolers' literacy, vocabulary, and math skills. *Developmental Psychology, 43*, 947–959. doi:10.1037/0012-1649.43.4.947

McClelland, M. M., Morrison, F. J., & Holmes, D. L. (2000). Children at risk for early academic problems: The role of learning-related social skills. *Early Childhood Research Quarterly, 15*, 307–329. doi:10.1016/S0885-2006(00)00069-7

Meador, D. N., Turner, K. A., Lipsey, M. W., & Farran, D. C. (2013). *Administering measures from the PRI learning-related cognitive self-regulation study*. Nashville, TN: Vanderbilt University, Peabody Research Institute.

Miller, M. R., Giesbrecht, G. F., Müller, U., McInerney, R. J., & Kerns, K. A. (2012). A latent variable approach to determining the structure of executive function in preschool children. *Journal of Cognition and Development, 13*, 395–423. doi:10.1080/15248372.2011.585478

Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., & Howert, A. (2000). The unity and diversity of executive functions and their contributions to complex "frontal lobe" tasks: A latent variable analysis. *Cognitive Psychology, 41*, 49–100. doi:10.1006/cogp.1999.0734

Moffitt, T. E., Arseneault, L., Belsky, D., Dickson, N., Hancox, R. J., Harrington, H., . . . Caspi, A. (2011). A gradient of childhood self-control predicts health, wealth, and public safety. *PNAS Proceedings of the National Academy of Sciences of the United States of America, 108*, 2693–2698. doi:10.1073/pnas.1010076108

Müller, U., Kerns, K. A., & Konkin, K. (2012). Test-retest reliability and practice effects of executive function tasks in preschool children. *The Clinical Neuropsychologist, 26*, 271–287. doi:10.1080/13854046.2011.645558

Mullola, S., Ravaja, N., Lipsanen, J., Alatupa, S., Hintsanen, M., Jokela, M., & Keltikangas-Jävinen, L. (2012). Gender differences in teachers' perceptions of students' temperament, educational competence, and teachability. *British Journal of Educational Psychology, 82*, 185–206. doi:10.1111/j.2044-8279.2010.02017.x

Nampijja, M., Apule, B., Lule, S., Akurut, H., Muhangi, L., Elliott, A. M., & Alcock, K. J. (2010). Adaptation of Western measures of cognition for assessing 5-year-old semi-urban Ugandan children. *British Journal of Educational Psychology, 80*, 15–30. doi:10.1348/000709909X460600

Osborne, A. F., Butler, N. R., & Morris, A. C. (1984). *The social life of Britain's five year olds: A report of the Child Health and Education Study*. London, England: Routledge and Kegan Paul.

Passolunghi, M. C., & Cornoldi, C. (2008). Working memory failures in children with arithmetical difficulties. *Child Neuropsychology, 14*, 387–400. doi:10.1080/09297040701566662

Ponitz, C. C., McClelland, M. M., Matthews, J. S., & Morrison, F. J. (2009). A structured observation of behavioral regulation and its contributions to kindergarten outcomes. *Developmental Psychology, 45*, 605–619. doi:10.1037/a0015365

Potter, D., Mashburn, A., & Grissmer, D. (2013). The family, neuroscience, and academic skills: An interdisciplinary account of social class gaps in children's test scores. *Social Science Research, 42*, 446–464. doi:10.1016/j.ssresearch.2012.09.009

Raghubar, K. P., Barnes, M. A., & Hecht, S. A. (2010). Working memory and mathematics: A review of developmental, individual difference, and cognitive approaches. *Learning and Individual Differences, 20*, 110–122. doi:10.1016/j.lindif.2009.10.005

Raver, C. C., Carter, J. S., McCoy, D. C., Roy, A., Ursache, A., & Friedman, A. (2012). Testing models of children's self-regulation within educational contexts: Implications for measurement. *Advances in Child Development and Behavior, 42*, 245–270. doi:10.1016/B978-0-12-394388-0.00007-1

Rothbart, M. K., Ahadi, S. A., Hershey, K. L., & Fisher, P. (2001). Investigations of temperament at three to seven years: The Children's Behavior Questionnaire. *Child Development, 72*, 1394–1408. doi:10.1111/1467-8624.00355

Séguin, J. R., Nagin, D., Assaad, J.-M., & Tremblay, R. E. (2004). Cognitive-neuropsychological function in chronic physical aggression and hyperactivity. *Journal of Abnormal Psychology, 113*, 603–613. doi:10.1037/0021-843X.113.4.603

Speece, D. L., & Cooper, D. H. (1990). Ontogeny of school failure: Classification of first-grade children. *American Educational Research Journal, 27*, 119–140. doi:10.3102/00028312027001119

St. Clair-Thompson, H. L., & Gathercole, S. E. (2006). Executive functions and achievements in school: Shifting, updating, inhibition, and working memory. *The Quarterly Journal of Experimental Psychology, 59*, 745–759. doi:10.1080/17470210500162854

Toplak, M., West, R., & Stanovich, K. (2013). Practitioner review: Do performance-based measures and ratings of executive function assess the

same construct? *Journal of Child Psychology and Psychiatry*, 54, 131–143. doi:10.1111/jcpp.12001_2012

Wanless, S. B., McClelland, M. M., Acock, A. C., Chen, F. M., & Chen, J. L. (2011). Behavioral regulation and early academic achievement in Taiwan. *Early Education and Development*, 22, 1–28. doi:10.1080/10409280903493306

Welsh, J. A., Nix, R. L., Blair, C., Bierman, K. L., & Nelson, K. E. (2010). The development of cognitive skills and gains in academic school readiness for children from low-income families. *Journal of Educational Psychology*, 102, 43–53. doi:10.1037/a0016738

Wiebe, S. A., Espy, K. A., & Charak, D. (2008). Using confirmatory factor analysis to understand executive control in preschool children: I. Latent structure. *Developmental Psychology*, 44, 575–587. doi:10.1037/0012-1649.44.2.575

Woodcock, R. W., McGrew, K. S., & Mather, N. (2001). *Woodcock-Johnson III Tests of Achievement*. Rolling Meadows, IL: Riverside.

Zelazo, P. D. (2006). The Dimensional Change Card Sort (DCCS): A method of assessing executive function in children. *Nature Protocols*, 1, 297–301. doi:10.1038/nprot.2006.46

Appendix

Multilevel Regression Equation for Model 4 (See Model 4 in Table 4)

All multilevel models were run in IBM SPSS (Version 20 Mixed Models) using restricted maximum-likelihood estimation. Provided below is a sample model equation for the EF direct assessments and teacher reports entered simultaneously to predict language outcomes. In the Level 1 equation, the posttest language score (Y) for a child (i) who is in classroom (j) situated in school (k) and system (l) is a function of the intercept of the mean language score (β_{0jkl}) and the fixed effects associated with a vector of demographic covariates, including pretest language score, interval of time elapsed between pretest and posttest, age at pretest, gender, and individualized education plan status ($\Sigma\beta_{1jkl}$). A child's posttest language score is also a function of the child's teacher-reported EF skills (β_{2jkl}), the individual child direct assessments of EF skills (β_{3jkl} , β_{4jkl} , β_{5jkl} , β_{6jkl} , β_{7jkl} , and β_{8jkl}), and the Level 1 random effect of the mean language score for children in each classroom (ε_{ijkl}). In the Level 2 equation, the intercept is a function of the classroom mean language score (γ_{00kl}), the fixed effects of the Level 1 predictors ($\gamma_{1000} \dots \gamma_{8000}$), and the Level 2 random effect associated with the intercept (η_{0jkl}). At Level 3, the intercept is a function of the school mean language score (π_{000l}), the experimental condition assignment of the school (π_{001l}), and the Level 3 random effect associated with the intercept (ξ_{000l}). Last at Level 4, the intercept is a function of the system mean language score (μ_{0000}), the Level 3 experimental condition (μ_{0010}), and the Level 4 random effect associated with the intercept (ω_{000l}).

Level 1 (child level):

$$Y_{ijkl} = \beta_{0jkl} + \sum \beta_{1jkl}(\text{Covariates}) + \beta_{2jkl}(\text{TEACHER REPORT}) + \beta_{3jkl}(\text{CORSI FORWARD})$$

$$+ \beta_{4jkl}(\text{CORSI BACKWARD}) + \beta_{5jkl}(\text{DCCS}) + \beta_{6jkl}(\text{HTKS}) + \beta_{7jkl}(\text{PEG TAPPING}) + \beta_{8jkl}(\text{COPY DESIGN}) + \varepsilon_{ijkl}$$

Level 2 (classroom level):

$$\begin{aligned} \beta_{0jkl} &= \gamma_{00kl} + \eta_{0jkl} \\ \beta_{1jkl} &= \gamma_{1000} \\ \beta_{2jkl} &= \gamma_{2000} \\ \beta_{3jkl} &= \gamma_{3000} \\ \beta_{4jkl} &= \gamma_{4000} \\ \beta_{5jkl} &= \gamma_{5000} \\ \beta_{6jkl} &= \gamma_{6000} \\ \beta_{7jkl} &= \gamma_{7000} \\ \beta_{8jkl} &= \gamma_{8000} \end{aligned}$$

Level 3 (school level):

$$\gamma_{00kl} = \pi_{000l} + \pi_{001l}(\text{CONDITION}) + \xi_{000l}$$

Level 4 (system level):

$$\begin{aligned} \pi_{000l} + \mu_{0000} + \omega_{000l} \\ \pi_{001l} = \mu_{0010} \end{aligned}$$

Received January 28, 2013

Revision received May 21, 2014

Accepted June 11, 2014 ■

The Effect of Training and Consultation Condition on Teachers' Self-Reported Likelihood of Adoption of a Daily Report Card

Alex S. Holdaway and Julie Sarno Owens
Ohio University

Using a within-subjects design and validated vignettes, this study examined the relative effects of four training and consultation conditions (i.e., consultation with key opinion leaders, consultation with observation and performance feedback, consultation with motivational interviewing, and professional development-as-usual) on teachers' ($N = 157$) self-reported ratings and rankings of the likelihood of adoption of a daily report card intervention for students with disruptive behaviors. The consultation with key opinion leaders condition produced significantly higher ratings of the likelihood of reported adoption than did the consultation with motivational interviewing or professional development-as-usual conditions, and was ranked higher than all other conditions. Professional development-as-usual was rated and ranked significantly lower than all other conditions. Teacher factors, including teacher experience and teacher burnout, were evaluated as predictors of adoption ratings. Implications and recommendations regarding the use of training and consultation conditions in research and practice are discussed.

Keywords: teacher consultation, teacher professional development, classroom intervention, daily report card, intervention adoption

Millions of dollars and hours are spent by school systems each year so that teachers are theoretically prepared and proficient to use up-to-date, evidence-based instruction, technology, and classroom management skills. Research on best practices for professional development (PD) suggest that programs that successfully produce change in teacher behavior are those that are of sufficient duration to learn and test new skills, include opportunities for active learning such as modeling and demonstration, and provide ongoing consultation¹ to support the specific skills being taught, and have content tailored to the teacher's specific situation (i.e., grade level, occupational responsibilities) (Darling-Hammond, Chung Wei, Andree, Richardson, & Orphanos, 2009; Yoon et al., 2007). Although evidence-based guidelines for PD are available, research suggests that the majority of PD programming offered to teachers is delivered in a brief workshop format with little or no opportunity for skill acquisition or ongoing follow-up (Darling-Hammond et al., 2009; Yoon et al., 2007) and is limited in changing teacher practices and behavior. Indeed, the results of one survey indicated that when asked to consider the last 3 years of PD experiences, less than 25% of teachers reported that their PD experiences have had an impact on their instruction (Hudson, McMahon, & Overstreet, 2002).

National surveys indicate that few teachers report feeling adequately trained to manage student disruptive behavior (National Comprehensive Center for Teacher Quality, 2012; The New Teacher Project, 2013), and elementary school teachers rank classroom management as their second greatest area of need for PD, behind only instructional skills (Coalition for Psychology in Schools and Education, 2006). In one study of over 7,000 educators, less than 30% perceived their training in student discipline and management as "useful" or "highly useful" (Darling-Hammond et al., 2009). Thus, not only are few teachers exposed to PD training in classroom management, but the PD literature suggests that the most frequently used methods are likely insufficient to change actual teacher practices (Darling-Hammond et al., 2009; National Council on Teacher Quality, 2014; National Research Council, 2001). Given these limitations, it may be unrealistic to expect teachers to adopt evidence-based behavior management interventions without supports that differ substantially from PD-as-usual (i.e., a workshop with no follow-up support). Because teacher-reported intentions to adopt a practice are moderately correlated with actual behavior (Armitage & Conner, 2001) and teacher perceptions of interventions have been shown to be positively associated with sustained use of the intervention (Baker, Kupersmidt, Voegler-Lee, Arnold, & Willoughby, 2010), examining teacher-reported intentions to adopt an intervention, given specific types of training and consultation, may be an effective way for researchers and administrators invested in choosing intervention training and consultation supports to help forecast how

This article was published Online First July 28, 2014.
Alex S. Holdaway and Julie Sarno Owens, Center for Intervention Research in Schools, Ohio University.

Correspondence concerning this article should be addressed to Julie Sarno Owens, 243 Porter Hall, Department of Psychology, Ohio University, Athens, OH 45701. E-mail: owensj@ohio.edu

¹ We use the term *consultation* throughout this document to describe individual meetings between a teacher and another individual tasked with helping improve implementation of an intervention. However, it is important to note that the term *coaching* could also have been used, as clearly differentiated definitions of consultation and coaching have not been well articulated in the literature (Denton & Hasbrouck, 2009).

teachers will respond when provided with differing types of intervention supports.

Using a within-subjects design and validated vignettes, we examined the relative effects of three enhanced training and consultation conditions (hereafter referred to as TCCs) in comparison to a PD-as-usual condition on teachers' perceptions of the likelihood of adopting an evidence-based behavior management intervention. Specifically, the study addresses the following questions: (a) Do enhanced TCCs produce higher ratings and rankings of likelihood of adoption of an evidence-based classroom intervention as compared with PD-as-usual? (b) Do specific, teacher-level factors predict ratings of adoption in the context of each TCC? Below, a critical review of promising TCCs as well as factors that may impact teachers' likelihood of intervention adoption is provided.

Teacher Adoption of Interventions

In the field of implementation science, the study of the translation of research to practice, several frameworks have been proposed to convey the phases that individuals and organizations go through when faced with the decision to adopt and implement a new intervention (e.g., Aarons, Hurlburt, & Horwitz, 2011; Fixsen, Blase, Naoom, & Wallace, 2009). Across the various frameworks, adoption represents the first phase and is followed by the implementation and sustainment phases. Although exposure to and experimentation with a given intervention represents the adoption phase, adoption is often conceptualized as a "one-time event" that an individual or organization makes, that is, a decision to adopt or reject the intervention (Aarons et al., 2011, p. 9). *Implementation* refers to the active use of an intervention once it has been adopted. *Acceptability* refers to whether an intervention is viewed by stakeholders (e.g., teachers) as appropriate for the problem, fair, and reasonable (Kazdin, 1980). Theoretically, interventions with higher acceptability ratings are more likely to be adopted, implemented with high integrity, and sustained than interventions with lower acceptability ratings. Although intervention adoption and implementation are complex processes, these definitions offer a useful delineation for the purpose of this study.

The authors of the implementation models have also identified factors theorized to enhance or interfere with adoption and implementation, including characteristics of the intervention, the teacher (or implementer), and the school building or district. Although the three phases can overlap and are not limited to a linear sequence (Fixsen et al., 2009), adoption may be particularly important to evaluate for two reasons: (a) The theoretical models suggest that by understanding factors that positively influence adoption, school systems can increase efficiencies in the dissemination of intervention to children in need and (b) such enhanced efficiencies may reduce costs and wasted resources, and potentially increase the likelihood of continued implementation and sustainment.

However, though theoretical models have emerged that describe the potential importance of the adoption process, few studies have examined the effect of PD supports on perceptions of the "adoptability" of the intervention. In one example, researchers examining teacher participation in a universal prevention program designed to reduce problem behavior and enhance school readiness in preschoolers found that teachers' concerns *prior to* any training or implementation were negatively associated with actual participation (i.e., implementation of the intervention) throughout the study

(Baker et al., 2010). Further, teacher characteristics, including perceptions of professional support and occupational satisfaction, were positively related to participation. Therefore, it is possible that by either (a) identifying or adapting an intervention such that the intervention results in fewer preadoption teacher concerns or (b) better identifying those teachers who may be in need of extra resources or supports to adopt and implement interventions, teacher adoption and implementation rates might be enhanced, significantly reduce wasted resources and costs, and increase efficiencies in extending the reach of interventions to students. However, as noted by implementation science theorists, empirical study of the posited conceptual models is in its infancy and, to date, has largely been focused on the implementation phase rather than the adoption phase (Fixsen et al., 2009). Because the adoption phase sets the stage for the remaining processes, research on the adoption of interventions by teachers could have a significant impact on the development of, and costs associated with, teacher PD. Below, enhanced models of teacher PD that may have a positive effect on teacher's adoption of evidence-based behavior management interventions are reviewed.

Consultation With Key Opinion Leaders

One proposed mechanism for increasing adoption of interventions is the use of key opinion leaders (Atkins et al., 2008). Key opinion leader procedures emphasize the use of respected peer teachers as intervention advocates and consultants to disseminate intervention components. This approach theoretically capitalizes on the advantages in contact, respect, and trust that a key opinion leader teacher possesses over an external consultant to improve intervention adoption and implementation. In published trials, key opinion leaders have held a role similar to an informal consultant. Specifically, key opinion leaders speak with peers about which techniques have worked in their classroom, discuss barriers and help problem-solve implementation challenges, but do not engage in specific observations or skill-based practices (Atkins et al., 2008). Further, key opinion leader support has been conceptualized as consultation on an as-needed basis, as opposed to regularly scheduled consultation meetings. In the only published reports of key opinion leader influence on teacher adoption of a classroom intervention, researchers found that teachers self-reported greater adoption and implementation of intervention strategies when working jointly with key opinion leaders and mental health consultants than when working with mental health consultant support alone. However, this initial advantage weakened over time such that the groups did not differ by the end of the trial's second year (Atkins et al., 2008). The diminishing effect of key opinion leader influence over time may indicate that key opinion leader influence is a useful tool to promote initial adoption of an intervention, but may be relatively less effective as a means of ongoing implementation support.

Consultation With Observation and Performance Feedback

Some teachers may elect not to adopt an intervention due to limited skill development in behavior management, experience with the intervention, or the opportunity to receive guidance about skill implementation. Performance Feedback typically involves

observations of the teacher's implementation, followed by a review of data from the observation in graphic and/or written form, highlighting the teacher's strengths and areas for improvement with the intervention components (Coddng et al., 2005). Performance feedback consultation meetings are scheduled on a regular basis and occur over several weeks or months, until high-quality implementation is achieved. Although some inconsistencies exist, studies provide compelling data that performance feedback produces higher levels of integrity to intervention procedures relative to baseline conditions or alternative strategies, that integrity declines precipitously in the absence of performance feedback, and that performance feedback can be delivered in a manner that is acceptable to teachers (e.g., Coddng et al., 2005; Noell et al., 1997). Compared with other options (e.g., key opinion leader and motivational interviewing), performance feedback has a comparatively strong research base, with a large number of single-case studies and a handful of randomized controlled trials documenting its effectiveness in increasing implementation integrity (Solomon, Klein, & Politylo, 2012). Although these findings are encouraging, they yield little information about the likelihood that a teacher may elect to *adopt* an intervention for use in his or her classroom if intervention support includes performance feedback. Nonetheless, if teachers know that they will receive such ongoing support, it may address preintervention concerns that affect their decision to adopt an intervention.

Consultation With Motivational Interviewing

A third approach that is effective in improving adoption behaviors in a number of health and mental health contexts is motivational interviewing (Miller & Rollnick, 2013). Although motivational interviewing has primarily been studied as an approach to improve the motivation of substance users and clinical populations to change health behaviors, consultation with motivational interviewing has recently been adapted as a tool to increase teacher adoption and integrity to intervention procedures (Frey et al., 2013; Gueldner & Merrell, 2011; Reinke et al., 2012). Though differences exist in conceptualizations of motivational interviewing in the school context, consultation with motivational interviewing typically includes having the teacher meet one-on-one with a consultant to (a) explore the teacher's values, (b) assess current practices via an interview and/or classroom observation, and (c) provide feedback on teacher's current practices with exploration of teacher perceptions of evidence-based strategies that may address areas of concern (see Frey et al., 2013, for a helpful guide). Throughout this short-term process, consultants attempt to enhance initial motivation to adopt the intervention by connecting teacher values with the intervention (e.g., "I know you've told me before that you really value independence in your students. How might this intervention impact their independence?") and eliciting "change talk" by specifically highlighting the discrepancy between the teacher's values and the status quo (e.g., "Achievement is important to you. You also have mentioned to me that you have a number of students with behavior issues that aren't doing well academically. I wonder if this intervention could help them?"). Another important distinction is that, as compared with key opinion leaders and performance feedback, motivational interviewing does not necessarily include ongoing support throughout the year, although such procedures could be used in combination (e.g.,

Reinke et al., 2012). Though studies that use motivational interviewing as a theoretical guide for teacher consultation are growing (e.g., Frey et al., 2013; Gueldner & Merrell, 2011; Reinke, Herman, & Sprick, 2011), there is still limited evidence of the impact of motivational interviewing on teacher *adoption* of interventions, specifically. As motivational interviewing is intended to directly affect teacher engagement and "buy-in," data as to how teacher's perceive consultation with motivational interviewing as compared with other TCCs are a critical step toward establishing its utility in enhancing intervention adoption.

Teacher Factors Associated With Adoption

Because there may not be a uniform preference for a single TCC across all teachers, identifying factors that predict adoption could have implications for differential consultation programming. Namely, consultation could be effectively tailored to the needs and characteristics of different subsets of teachers within a building or district. Predictors of adoption could also help to identify those teachers who may need additional encouragement or individualized attention if they are identified as less likely to adopt an intervention. Thus, factors that have been shown to be associated with the adoption and implementation of interventions were also examined as potential predictors of adoption.

Teacher Experience

New teachers often feel unprepared for the classroom behavior challenges that arise when educating students with disruptive behavior and consistently rate classroom behavior as a top reason for leaving the profession (Ingersoll, 2001). In a sample of over 2,000 educators, 52% of first-year teachers ranked classroom management as their number one PD need, with only 10% of teachers with over 10 years of experience ranking it as their number one PD need (Coalition for Psychology in Schools and Education, 2006). Although this finding suggests differences between teachers' priorities related to levels of experience, the impact of *specific types* of TCC approaches on teachers with differing levels of experience is unclear. The various components of teacher training and consultation programs such as working with peers (key opinion leader) or receiving feedback from outside observers (performance feedback) may be received differently by teachers who have a depth of experience and are perceived as school leaders, as compared with new or less experienced teachers.

Teacher Burnout

According to Maslach and colleagues (1996), burnout is characterized by emotional exhaustion, disengagement, and a low sense of personal accomplishment. Studies of teacher burnout have revealed that student misbehavior is a significant predictor of teacher burnout (Hastings & Bham, 2003). Further, teachers with higher levels of burnout were found to endorse more negative attitudes about implementing a new academic program than colleagues with lower levels of burnout (Evers et al., 2002). Thus, it is possible that this finding would generalize to teachers' attitudes about implementing a new intervention targeting student misbehavior.

Self-Efficacy

Teacher self-efficacy is a teacher's belief that he or she can perform a classroom procedure with a high degree of competence. A vignette-based study has shown that teachers high in self-efficacy rated consultation as being more effective and acceptable than teachers low in self-efficacy (DeForest & Hughes, 1992). Studies examining the relationship between self-efficacy and actual intervention adoption document higher rates of adoption by teachers with higher self-efficacy than by teachers with lower self-efficacy (Rimm-Kaufman & Sawyer, 2004). This study expands the exploration of teacher self-efficacy into the TCC domain.

Use of Behavioral and Instructional Strategies

Many interventions for children with disruptive behaviors rely on behavioral principles such as contingency management and reinforcement. If teachers are already using behavioral principles and reinforcement in their classroom, it may be that teachers will be more likely to adopt a behaviorally based intervention, as it aligns with their current practices. As such, a measure of teachers' use of behavioral and intervention strategies was included.

Principal Support

Principals are commonly the individuals tasked with being the "gatekeepers" for new programs introduced at the school (Hallinger & Heck, 1996), with substantial influence over time spent, resource allocation, and incentives for adoption and implementation. One study shows that in the dissemination of an evidence-based intervention for youth at risk for delinquency, positive intervention outcomes occurred only in those schools that had both high principal support and a high degree of classroom implementation by teachers (Kam, Greenberg, & Walls, 2003). However, no studies have specifically examined the effects of principal support on teachers' perceptions of intervention adoption for students with disruptive behavior.

Interventions for Students With Disruptive Behavior

The above literature review highlights the TCCs that may enhance adoption of an intervention relative to PD-as-usual and factors that may impact teachers' likelihood of intervention adoption. Now attention is turned to a specific evidence-based behavioral intervention that teachers could adopt, namely, the daily report card (DRC). The DRC is a well-established intervention with strong empirical support for effectiveness in both general and special education settings (Kelley, 1990; Owens et al., 2012; Vannest, Davis, Davis, Mason, & Burke, 2010). The DRC is a tool to monitor and modify clearly defined target behaviors (e.g., interruptions, work completion) by setting daily goals for the student (e.g., seven or fewer interruptions per day), providing feedback to the student, and providing rewards for attaining daily goals (Kelley, 1990). Despite teachers' reported acceptability of the DRC (e.g., Girio & Owens, 2009), studies indicate that, outside of research-based programs, such interventions are underused (Martinussen, Tannock, & Chaban, 2011). By examining teacher preferences for, and perceptions of, enhanced TCCs, a better understanding of the types of PD that may lead to higher rates of

adoption and use of strategies that effectively address disruptive student behavior may be achieved. Given that educating students with disruptive behavior has been linked to increased teacher stress, negative teacher-student interactions, and has been identified as a major contributor to teacher turnover and job dissatisfaction (Brouwers & Tomic, 2000; Greene, Besztercze, Katzenstein, Park, & Goring, 2002; Ingersoll, 2001), understanding how to facilitate teachers' adoption and use of an intervention that improves the behavior of students with disruptive behavior, while reducing teacher stress and school expenditures, is an important priority.

The Current Study

This study addresses limitations in the teacher PD literature by (a) simultaneously examining the effects of multiple TCCs on reported adoption likelihood ratings, (b) examining the effects of TCCs on adoption rather than implementation or sustainability, and (c) concurrently examining multiple teacher factors to identify predictors of adoption likelihood ratings. These advancements allow researchers and school-based administrators to compare teacher perceptions of each TCC, identify unique predictors of reported adoption decisions, and disentangle the effects of the intervention itself from the TCC provided.²

The first aim was to answer the questions: Do enhanced TCCs produce higher ratings and rankings of likelihood of adoption of an evidence-based classroom intervention as compared with PD-as-usual? If so, which TCC do teachers prefer and rate as most likely to result in intervention adoption? Given teacher reports of the current state of PD (Hudson et al., 2002), we hypothesized that key opinion leader, performance feedback, and motivational interviewing TCCs would be preferable to the PD-as-usual condition. The second aim was to answer the question: Do specific, teacher-level factors predict ratings of adoption in the context of each TCC? Given that multiple factors have not previously been examined simultaneously, hypotheses about the relative predictive utility of each factor were not made.

Method

Participants

An a priori power analysis, using G*Power, version 3.1.2 (Faul, Erdfelder, Lang, & Buchner, 2007), was performed to determine an appropriate sample size for the analyses, with priority given to the repeated measures analysis of variance (ANOVA) for the first aim. Input parameters were conservatively selected (effect size = .1, $\alpha = .05$, $1 - \beta = .8$) so that small between-condition differences could be detected. It was determined that 138 individuals would be adequate to achieve the desired power for the analyses associated with the first aim.

Participants were 157 teachers (87.9% female; 96.8% Caucasian; 61.1% with a master's degree; 22.6% with a special education certification) from eight schools in three school districts in

² We could have examined various combinations of the TCCs, as teachers may have access to more than one at a time; however, understanding the impact of individual TCCs was considered to be the most parsimonious and prudent at this stage of the research.

southeastern Ohio, selected on the basis of geographic proximity and administrator amenability to study procedures (i.e., a sample of convenience). Years of experienced ranged from 1 to 43 ($M = 17.83$, $SD = 14.71$), and the average teacher age was 43.24 ($SD = 11.21$). Participating schools serve approximately 2,750 students (97.4% Caucasian) in grades pre-K through Grade 6, of which 54%–78% of students were eligible for free and reduced-price lunches (U.S. Department of Education, 2009) and 15.2% had a reported disability. For six schools, study measures were collected prior to an in-service training in August 2011. For the remaining two schools, study measures were collected in a group format during faculty meetings in January and February 2012. The overall teacher response rate was 85.8%. Individual school response rates ranged from 83.7% to 87.7%.

Measures

Demographic data and principal support. Teachers reported years of teaching experience and perceptions of principal support. Using a 4-point scale (ranging from 1 *not supportive* to 4 *very supportive*), teachers rated how supportive the principal was of (a) general use of classroom interventions for children with disruptive behavior and (b) the teacher, personally, using classroom interventions for children with disruptive behavior.

Use of behavioral and instructional approaches. The *Instructional and Behavior Management Approaches Survey* (Martinussen et al., 2011) is a 39-item scale that measures teachers' self-reported frequency of use of instructional and behavior management approaches. Items are rated on a 5-point scale ranging from 1 (*rarely*) to 5 (*most of the time*). The survey has two subscales: Behavior Management Techniques (19 items) and Instructional Approaches (20 items). Because the survey authors found the two subscales to be highly correlated ($r = .74$), and previous studies using similar measures have not differentiated the two subscales (e.g., Fabiano et al., 2002), a total scale score was used in this study. Three items were removed (verbal reprimands, remove student from class, and lowering expectations for work) due to corrected item-total correlations less than .2. The internal consistency coefficient for the 36-item scale was .93. Possible scores range from 36 to 180.

Teacher self-efficacy. The *Ohio State Teacher Efficacy Scale* (Tschanen-Moran & Hoy, 2001) is a 12-item scale consisting of three subscales: Efficacy for Instructional Strategies, Efficacy for Classroom Management, and Efficacy for Student Engagement. Subscale scores range from 4 to 36; higher scores indicate higher efficacy. The measure has adequate convergent validity with measures of teacher self-efficacy (Tschanen-Moran & Hoy, 2001) and has been found to have a three-factor structure in exploratory and confirmatory analyses (Tschanen-Moran & Hoy, 2001). Internal consistency coefficients in the current study were .91 (total scale), .81 (instructional strategies subscale), .89 (classroom management subscale), and .84 (student engagement subscale).

Teacher burnout. The *Maslach Burnout Inventory-Educators Survey* (MBI-ES; Maslach & Jackson, 1986) is a 22-item self-report scale that assesses teacher burnout. The measure yields a total score and three subscales: Emotional Exhaustion, Depersonalization, and Sense of Personal Accomplishment. Items are rated on a scale ranging from 0 (*never*) to 6 (*every day*). Subscale scores range from 0 to 54. The MBI-ES is one of the most

often used measures of burnout in the field of education and has been found to have a consistent three-factor structure across investigations of educator burnout (see Worley, Vassar, Wheeler, & Barnes, 2008, for a meta-analysis). Studies distinguished burnout from job dissatisfaction and established burnout as a distinct construct (Leiter & Durup, 1994). Internal reliability coefficients were .90 for Emotional Exhaustion, .76 for Depersonalization, and .76 for Sense of Personal Accomplishment (Iwanicki & Schwab, 1981). In the current sample, alpha coefficients were .90, .69, and .79, for Emotional Exhaustion, Depersonalization, and Sense of Personal Accomplishment, respectively.

Likelihood of intervention adoption. The *Intervention Support Questionnaire* (ISQ) is a self-report questionnaire designed for the current study to assess the likelihood of DRC adoption when teachers are provided with different TCCs. The ISQ includes vignette descriptions of (a) a child demonstrating moderate levels of attention-deficit/hyperactivity disorder symptom severity who is moderately disruptive in the classroom; (b) a teacher-led individualized intervention (i.e., DRC) that was appropriate and feasible to address the child's difficulties; and (c) descriptions of TCCs that reflect consultation with a key opinion leader, consultation with performance feedback, consultation with motivational interviewing and PD-as-usual procedures (see ISQ construction and validation information in the Appendix). After reading each TCC description, participants reported the likelihood of adopting the intervention given the TCC provided using a scale that ranged from 0 (*There is no chance I will use this intervention*) to 100 (*I will definitely use this intervention*), with 10-point increments. The ISQ also provides space to write a narrative description of what aspects of the TCC description influenced the respondent's likelihood of adoption ratings (analysis of these data are in progress). After reading all four TCCs, respondents also were asked to rank order the TCCs from 1 (*most likely to result in intervention adoption*) to 4 (*least likely to result in intervention adoption*) in a forced-choice format.

Data Collection Procedures

All procedures were reviewed and approved by the university Institutional Review Board in an expedited review. Participation was voluntary. Teachers received a token of appreciation with value less than \$10 for participation. Teachers were provided with a packet containing a consent form and all questionnaires. Within each packet, questionnaire order was counterbalanced. Further, using a within-subjects design, the order of vignettes was also counterbalanced within the ISQ to address potential ordering effects. Consent forms and verbal descriptions (for participants who completed forms in person) indicate that no identifiable results would be communicated with school administrators or principals and that the purpose for the study was to examine teachers' preferred intervention supports. Participants were instructed to complete all forms and return the packet to research staff. Packets were left at the schools for teachers unable to attend the data collection meetings and retrieved 1 week later (14.6% of the sample). No differences were found between teachers who completed the packets in person or independently, between teachers from different schools, or between teachers who completed measures in summer versus winter.

Results

Aim 1: Do Enhanced TCCs Produce Higher Reports of Likelihood of Intervention Adoption as Compared With PD-as-Usual?

Likelihood of adoption ratings. Teachers' self-reported likelihood of intervention adoption rates were examined using a one-way repeated measure ANOVA. TCC was a repeated measures variable because each teacher completed ratings and rankings for all four TCCs. Omnibus results determined that mean likelihood of adoption likelihood ratings differed significantly as a function of TCC, $F(3, 465) = 58.79, p < .001$. The average rates of likelihood of adoption for each condition are presented in Table 1.

Post hoc tests of adoption ratings using the Bonferroni correction and Morris and Deshon's (2002) equation for within-subject effect sizes, controlling for dependence, revealed that the key opinion leader condition received the highest mean likelihood of adoption rating among the TCCs ($M = 79.58$) and was rated significantly higher than the motivational interviewing ($M = 68.75$; $ES = .47, p < .001$) and PD-as-usual conditions ($M = 56.4$; $ES = .93, p < .001$), but not the performance feedback condition ($M = 75.42$; $ES = .21, p = .070$). The performance feedback condition was rated significantly higher than the motivational interviewing ($ES = .30, p < .01$) and PD-as-usual condition ($ES = .75, p < .001$). Motivational interviewing was rated significantly higher than the PD-as-usual condition ($ES = .52, p < .001$). Thus, teachers reported a greater likelihood of adoption of a DRC intervention when presented with the key opinion leader or performance feedback conditions than if presented with motivational interviewing or PD-as-usual, and a greater likelihood of adoption if presented with motivational interviewing than if presented with PD-as-usual. As hypothesized, the key opinion leader, performance feedback, and motivational interviewing TCCs produced higher ratings than the PD-as-usual condition (see Table 1).

Likelihood of adoption rankings. The average adoption rankings for each condition are presented in Table 1. To examine teachers' likelihood of adoption rankings, a Wilcoxon signed-rank test was conducted (see Table 1). The key opinion leader condition was ranked as the most preferred strategy, significantly more often than performance feedback ($Z = -3.558, p < .001$), motivational

interviewing ($Z = -6.821, p < .001$), and PD-as-usual ($Z = -7.837, p < .001$). The performance feedback condition was ranked second highest and significantly more often than the motivational interviewing condition ($Z = -3.295, p < .001$) and PD-as-usual condition ($Z = -5.986, p < .001$). Finally, the motivational interviewing condition was ranked as the third most preferred strategy, significantly more often than the PD-as-usual condition ($Z = -3.166, p < .001$). Thus, when presented with a forced-choice ranking format, teachers ranked the key opinion leader condition as most likely to result in intervention adoption, followed by the performance feedback, motivational interviewing, and PD-as-usual conditions. The percentages of each rank that each TCC received can be found in Table 2.

Aim 2: Do Teacher-Level Factors Predict Ratings of Adoption in the Context of Each TCC?

Ten predictors of likelihood of adoption ratings were examined simultaneously in a series of four linear regressions; one regression model per TCC. Each regression included an examination of variance inflation factors (VIFs) to check for collinearity. No variables needed removal, as all VIFs fell below a VIF value of 5. See Table 3 for descriptive information for predictor variables, Table 4 for correlations between predictors and TCC ratings, and Table 5 for regression results. Correlations between predictors are presented in Table 6.

Consultation with key opinion leaders. The total regression model for the key opinion leader condition was not significant ($R^2 = .092$), $F(10, 428.88) = 1.32, p = .225$. Number of years employed as a full-time teacher ($\beta = -.205$), $t(151) = -2.32, p = .022$, was the only significant predictor of key opinion leader ratings. When key opinion leader consultation was offered, teachers who have less experience provided higher likelihood of adoption ratings than teachers who have more experience. A follow-up median split analyses indicated that teachers with less than or equal to 15.50 years of experience provided higher ratings of adoption with key opinion leader support ($M = 83.82, SD = 16.97$) than did teachers with more than 15.50 years of experience ($M = 75.20, SD = 19.42$), $t(150) = 2.91, p = .004$.

Performance feedback. The total regression model for the performance feedback condition was not significant ($R^2 = .116$), $F(10, 701.72) = 1.70, p = .087$. The Sense of Personal Accomplishment subscale from the MBI-ES ($\beta = .222$), $t(151) = 2.28, p = .024$, was the only significant predictor. When consultation with performance feedback was offered, teachers who had a higher sense of personal accomplishment (e.g., "I feel I'm positively influencing other people's lives through my work") reported higher likelihood of adoption ratings than teachers who had a lower sense of personal accomplishment. The Sense of Personal Accomplishment subscale of the MBI-ES yields three categories of personal accomplishment: low, moderate, and high. Using a one-way ANOVA, average likelihood of adoption rating differences were examined between those participants who fell into each category. Results indicated significant differences in mean performance feedback rankings between conditions, $F(2, 153) = 6.90, p = .001$. Mean rankings were highest for those high in personal accomplishment ($n = 90; M = 79.83, SD = 21.86$), followed by those with a moderate ($n = 46; M = 73.91, SD = 18.91$) and low

Table 1
Intervention Adoption Likelihood Ratings and Rankings by TCC

Support type	Mode	M	SD
Ratings			
Key opinion leader	90	79.39 _a	18.72
Performance feedback	90	75.25 _a	22.20
Motivational interviewing	80	68.63 _b	23.20
PD-as-usual	50	56.41 _c	24.68
Rankings			
Key opinion leader	1	1.78 _a	0.97
Performance feedback	2	2.28 _b	1.03
Motivational interviewing	3	2.74 _c	0.92
PD-as-usual	4	3.19 _d	1.03

Note. Higher ratings reflect higher likelihood of adoption; lower rankings reflect lower likelihood of adoption. Conditions with different subscripts are significantly different at the $p < .05$ level. TCC = training and consultation condition; PD = professional development.

Table 2
Distribution of Rankings for Each TCC

Support technique	First	Second	Third	Fourth
Key opinion leader	52.6%	25.0%	14.5%	7.9%
Performance feedback	27.0%	32.9%	25.0%	15.1%
Motivational interviewing	9.2%	30.4%	36.2%	23.7%
PD-as-usual	11.2%	11.8%	23.7%	53.3%

Note. TCC = training and consultation condition; PD = professional development.

sense of personal accomplishment ($n = 20$; $M = 61.00$, $SD = 21.74$).

Consultation with motivational interview. There were no significant predictors in the model for the motivational interviewing condition.

PD-as-usual. There were no significant predictors in the model for PD-as-usual.

Discussion

As hypothesized, both ratings and rankings of adoption likelihood suggest that teacher perceptions of rates of adoption of an intervention for students with disruptive behavior may be significantly enhanced if intervention training and support includes consultation with key opinion leaders, consultation with performance feedback, or consultation with motivational interviewing compared with PD-as-usual. Further, the highest likelihood of intervention adoption was reported when teachers were provided with descriptions of consultation with key opinion leader and performance feedback. In partial support of theories of adoption and implementation (Fixsen et al., 2009; Rogers, 2003), some teacher-level predictors of reported adoption were identified. Below, implications for future research and practice are discussed.

Consultation With Key Opinion Leaders

According to theories of innovation diffusion (Rogers, 2003), key opinion leaders are important determinants of the potential

spread of an innovation (i.e., intervention) throughout an environment. Namely, respected, socially central figures in the school environment are conceptualized as models of behavior for others in the social network, and considerable weight is given to their opinions and experience. Consistent with this theory, as well as other recent studies (Atkins et al., 2008; Cunningham et al., 2009), nearly 80% of teachers (see Table 2) reported the key opinion leader condition to be the first or second most likely TCC to result in DRC adoption, with a mean rating of 79.39 across all teachers. Though this figure likely overstates the number of teachers who would adopt the intervention if provided a real opportunity, this perception may be an important indicator of true behavior and may also relate to sustained implementation (i.e., Baker et al., 2010).

These data, and those of Atkins and colleagues (2008), provide substantial evidence that teachers trust the opinions of respected colleagues in their network and feel comfortable receiving consultation and support from them. Indeed, this is the likely natural mechanism through which many teachers obtain advice and recommendations about both instructional and behavioral interventions, particularly in underresourced schools that lack consultants or school mental health professionals. Given the access that teachers have to key opinion leaders and the natural transfer of information that occurs within this social network, training key opinion leaders in evidence-based programs (in a manner that aligns with evidence-based PD models; Darling-Hammond et al., 2009) may be an effective catalyst for adoption by others.

It is important to note that our data and that of Atkins and colleagues (2008) are based on teacher self-report rather than actual observed intervention adoption. Though there is evidence to suggest that teacher perceptions of interventions, prior to training, relate to teachers' sustained implementation of interventions, these findings must be replicated (Baker et al., 2010). If replicated, schools could develop procedures for identifying the key opinion leader teachers who are respected for their skills (e.g., in behavior management) and develop a process for keeping them current in evidence-based practices and disseminating this information throughout the network of teachers. It is also important to note that the characteristics ascribed to key opinion leaders (e.g., well-

Table 3
Descriptive Statistics for Support Condition Rating Predictors

Teacher factor	<i>M</i>	<i>SD</i>	Minimum	Maximum
Experience				
Years of teaching experience	16.03	9.78	1	43
Burnout				
MBI-ES Emotional Exhaustion	20.96	10.67	0	47
MBI-ES Depersonalization	4.43	4.45	0	20
MBI-ES Sense of Personal Accomplishment	38.58	6.48	14	48
Self-efficacy				
OSTES Student Engagement	6.99	1.16	4	9
OSTES Instructional Strategies	7.43	.99	4	9
OSTES Classroom Management	7.32	1.09	4	9
Classroom management				
IBMAS total score	123.52	21.46	60	170
Principal support				
Principal support in general	3.40	.82	1	6
Principal support of individual	3.53	.88	1	6

Note. MBI-ES = Maslach Burnout Inventory-Educators Survey; OSTES = Ohio State Teacher Efficacy Scale; IBMAS = Instructional and Behavior Management Approaches Survey.

Table 4
Correlations Between Teacher Factors and TCC Ratings

Teacher factors	PF	KOL	MI	PDAU
Experience				
Years of teaching experience	-.150	-.230**	.020	-.070
Burnout				
MBI-ES Emotional Exhaustion	-.220	-.065	-.139	-.202*
MBI-ES Depersonalization	-.159*	-.073	-.094	-.203*
MBI-ES Sense of Personal Accomplishment	.260**	.149	.160*	.208*
Self-efficacy				
OSTES Student Engagement	.184*	.078	.073	.168*
OSTES Instructional Management	.171*	.053	-.005	.083
OSTES Classroom Management	.154	.003	.015	.076
Classroom Management				
IBMAS total	.105	.115	.077	.103
Principal support				
Principal support of school	.044	-.006	-.071	.109
Principal support of individual	.017	-.049	-.028	-.013

Note. TCC = training and consultation condition; PF = performance feedback; KOL = key opinion leader; MI = motivational interviewing; PDAU = professional development-as-usual; MBI-ES = Maslach Burnout Inventory–Educators Survey; OSTES = Ohio State Teacher Efficacy Scale; IBMAS = Instructional and Behavior Management Approaches Survey.

* $p < .05$. ** $p < .001$.

respected, available, experienced, skilled) might also be found in other professionals. Future studies may benefit from teasing apart what aspects of key opinion leaders are most important for adoption and to what extent these characteristics appear to be differentially associated with the individual characteristics of the person versus those inherent to the position as a peer teacher. It is also important to note that although key opinion leader teachers may enhance initial adoption behavior, there is no evidence that their impact extends to enhancing implementation integrity once the intervention is adopted. Indeed, enhancing integrity may exceed the reach, occupational responsibility, and skill set of a key opinion leader. To date, only intensive observation, coaching, and

performance feedback has evidence for enhancing implementation integrity (Domitrovich et al., 2008).

Consultation With Performance Feedback

Given the substantial evidence for performance feedback in enhancing implementation integrity, it is encouraging to see that teachers perceive consultation with performance feedback as more likely to result in adoption than procedures that have not been shown to result in behavior change (i.e., PD-as-usual). To our knowledge, no study has previously reported on teachers' perceptions of performance feedback *before* teachers had been exposed to

Table 5
Summary of Regression Analyses for Variables Predicting TCC Ratings

Teacher factor	Performance feedback		Key opinion leader		Motivational interviewing		Professional development-as-usual	
	R^2	β	R^2	β	R^2	β	R^2	β
Regression model	.116		.092		.076		.099	
Experience								
Years of teaching experience		-.093		-.205*		.059		-.027
Burnout								
MBI-ES Emotional Exhaustion		-.092		.075		-.075		-.059
MBI-ES Depersonalization		.006		-.060		-.020		-.080
MBI-ES Sense of Personal Accomplishment		.222*		.147		.190		.151
Self-efficacy								
OSTES Student Engagement		.116		.105		.138		.153
OSTES Instructional Management		.019		-.011		-.130		-.093
OSTES Classroom Management		-.051		-.120		-.082		-.050
Classroom management								
IBMAS total		.020		.044		.020		.078
Principal support								
Principal support in general		.120		-.009		-.072		-.178
Principal support of individual		-.192		-.078		-.058		.070

Note. TCC = training and consultation condition; β = standardized beta; MBI-ES = Maslach Burnout Inventory–Educators Survey; OSTES = Ohio State Teacher Efficacy Scale; IBMAS = Instructional and Behavior Management Approaches Survey.

* $p < .05$.

Table 6
Correlations Between Teacher Factors

Variable	1	2	3	4	5	6	7	8	9
1. Years employed	—								
2. MBI-ES Emotional Exhaustion	.061	—							
3. MBI-ES Depersonalization	.025	.631**	—						
4. MBI-ES Sense of Personal Accomplishment	-.142	-.356**	-.451**	—					
5. OSTES Student Engagement	-.031	-.317**	-.368**	.353**	—				
6. OSTES Instructional Management	.031	-.287**	-.375**	.331**	.563**	—			
7. OSTES Classroom Management	.104	-.264**	-.328**	.325**	.611**	.631**	—		
8. IBMAS total	-.113	-.018	-.103	.393**	.260**	.311**	.230**	—	
9. Principal support in general	-.138	-.178*	-.143	.191*	.254**	.235**	.348**	.148	—
10. Principal support of individual	-.169*	-.236	-.248**	.171*	.245**	.182*	.283**	-.018	.771**

Note. MBI-ES = Maslach Burnout Inventory—Educators Survey; OSTES = Ohio State Teacher Efficacy Scale; IBMAS = Instructional and Behavior Management Approaches Survey.

* $p < .05$. ** $p < .001$.

observation and feedback procedures. Our data offer promise that consultation with performance feedback is viewed positively by teachers and may enhance initial adoption decisions as compared with motivational interviewing and PD-as-usual conditions. Further, combining findings from this study and previous studies of performance feedback, examining a combined consultation protocol that includes both performance feedback and key opinion leaders may be a fruitful avenue for developing a system that impacts both initial adoption decisions and sustained implementation of interventions. Future studies could experimentally manipulate the use of key opinion leader and performance feedback procedures alone and in combination and examine the potentially differing impacts on teachers' actual adoption of the DRC and the integrity with which it is implemented over time.

It is interesting to consider the characteristics that separate the key opinion leader and performance feedback conditions from the motivational interviewing and PD-as-usual conditions that were rated and ranked significantly lower in producing adoption likelihood. The key opinion leader and performance feedback conditions were the only conditions in which consultation was described as being provided throughout the entire year, potentially suggesting that teachers recognize the need and/or benefits of ongoing support, consultation, and feedback, and that such ongoing support is an important determinant of teacher adoption decisions. Though further studies deconstructing which aspects of each of the TCCs are associated with rates of adoption are needed, this hypothesis aligns with best practices for PD (Darling-Hammond et al., 2009; Yoon et al., 2007). Future research should also examine the congruency between teacher reports of adoption and actual adoption behavior, as this finding may bolster researcher and consumer confidence in offering performance feedback as a mechanism for enhancing intervention adoption and using performance feedback throughout the year to sustain high-quality intervention implementation.

Finally, it is important to note that even though there is substantial support for the use of performance feedback, there is not agreement on the dosage or length of performance feedback that may be necessary or sufficient to produce sustained behavior change. Several studies document that brief versions of performance feedback can produce positive short-term outcomes. However, few studies have examined the long-term maintenance of

improvements in the quality of implementation, and a precipitous decline in quality implementation has been observed once the performance feedback is removed in some studies (e.g., Noell et al., 1997). Thus, performance feedback across an extended period or intensive performance feedback with periodic "boosters" may be necessary for sustained implementation.

Consultation With Motivational Interviewing

Although motivational interviewing was rated and ranked as less likely to result in DRC adoption than key opinion leader or performance feedback TCCs, our data suggest that interventions with a motivational interviewing component may still be a fruitful area for continued study. Indeed, teacher report suggests that consultation with motivational interviewing is perceived as significantly more likely to result in adoption than PD-as-usual. Researchers have yet to develop a consensus on how best to incorporate motivational interviewing principles in school-based consultation, but a variety of adaptations have been proposed (Frey et al., 2013; Reinke et al., 2012), and studies are currently underway to examine the extent to which the use of motivational interviewing in consultation enhances intervention adoption and implementation (Owens & Coles, 2014; Reinke, Frey, Herman, & Thompson, 2014). Given that 10% of teachers ranked consultation with motivational interviewing as their first choice of support, this TCC may be particularly helpful for a small subset of teachers, particularly those who may be ambivalent about the intervention. Because none of the teacher-level predictors in this study had significant utility in identifying this subset of teachers, research with different constructs such as attitudes toward evidence-based practices or the consultation process is warranted. In addition, future research could examine what, if any, incremental effect motivational interviewing procedures have, above and beyond performance feedback and key opinion leader approaches.

PD-as-Usual

As hypothesized, the PD-as-usual condition received the lowest rankings and ratings of the likelihood of DRC adoption. Although this finding was expected, this outcome is concerning, given that this style of PD is the most common method of

training available to teachers (Darling-Hammond et al., 2009). With less than 25% of teachers in this study ranking PD-as-usual as either first or second most likely to result in intervention adoption (Table 2), it would appear that the standard training and consultation options are perceived as largely insufficient, and indeed may explain the relatively low rates of adoption of evidence-based classroom interventions for youth with disruptive behaviors (Martinussen et al., 2011).

Predictors of Adoption

Although teacher-level factors (i.e., experience and personal accomplishment) show some promise in identifying teachers who are in need of further support or "matching" to a specific TCC, teacher-level factors explained a small amount of variance in likelihood of adoption ratings. Specifically, across TCC conditions, only two significant predictors of likelihood of adoption ratings emerged. First, results indicated that a higher sense of personal accomplishment was linked to higher adoption likelihood ratings in the performance feedback condition. This finding may indicate that teachers who have a higher confidence in their ability to influence student outcomes are more comfortable being observed, receiving feedback about their performance, and are more welcoming of consultation than those who question their impact on student outcomes. Second, more years of teacher experience was predictive of lower adoption likelihood ratings in the key opinion leader condition. This finding could be explained by a number of potential phenomena. First, it is possible that the more senior teachers are key opinion leaders themselves. Thus, others turn to them for support and advice, but they do not have others in the network to whom they turn for support and advice. Second, more senior teachers may feel as if they have already accumulated the knowledge and skills necessary to effectively manage disruptive behavior; thus, the presence of a key opinion leader may be less influential in their adoption decision. In contrast, the presence of a key opinion leader and a supportive collegial network may be incrementally important for more junior teachers, as there is evidence that teacher induction programs and collaborative environments are associated with teacher decisions to remain in the field and occupational satisfaction in young teachers (Kardos et al., 2001; Smith & Ingersoll, 2004). Thus, consultation with a key opinion leader may be a particularly attractive and important training and consultation option for new or less experienced educators (Shernoff et al., 2011). However, because eight of 10 predictors in the model were not significant, and because the significant predictors explained only a small portion of the variance in adoption ratings, the practical meaning of this finding may be limited. In addition, examination of the predictive utility of other teacher-level factors (e.g., attitude variables) may be more fruitful.

Limitations and Future Directions

First, this study is limited by its use of an analog design. Although steps were taken to ensure that all descriptions validly represented the intended constructs, it is unknown how these findings generalize to actual adoption of classroom interventions. However, evidence of a moderate correlation between self-predictions and actual behavioral enactment of intentions (Armit-

age & Conner, 2001), and the negative relationship between teacher report of concerns and participation in an intervention (Baker et al., 2010), offer some confidence that our results generalize to actual behavior. Second, all questionnaires and vignettes were specific to the teacher level. As schools are nested within school systems, and are influenced by state and federal education policy, future research should explore how these multilevel factors influence the adoption decisions about interventions for students with disruptive behaviors.

Third, it may be considered a limitation that TCCs were examined in isolation, rather than considering combinations of TCCs, as a combination may result in the highest likelihood of teacher adoption. Because this was the first study to directly compare multiple TCCs, it was important to examine parsimonious models first, and evaluate our hypotheses before examining more complex combinations. Additionally, this study did not explicitly test the impact of verbal consultation as compared with a more active "coaching" model of consultation-focused skill development through live coaching (e.g., "bug in the ear") or role-plays. Fourth, measurement of teacher-level predictors was potentially limited by the limited psychometric strength of measurement instruments used. Lastly, our sample was largely homogenous with regard to race, ethnicity, gender, and school type (i.e., rural), limiting the generalizability of the findings. Future research should replicate the current study with diverse populations. Finally, though efforts were taken to validate TCC descriptions, it is possible that the writing style and attention given to specific aspects of each TCC influenced teacher response. For example, responses may have differed had more explicit descriptions of time spent in consultation or magnitude of expected behavior change been operationally defined.

Conclusion

Children with disruptive behaviors have a number of functional impairments in the school setting that contribute to poor academic, social, and behavioral outcomes; teacher distress; and costly services. Although evidence-based individualized programs that have the potential to effectively treat disruptive behavior are available (Pelham & Fabiano, 2008), these interventions are seldom adopted and used outside of research contexts (Martinussen et al., 2011). Thus, for evidence-based classroom interventions to have their intended impact, mechanisms for enhancing teachers' adoption and high-quality implementation of such interventions must be identified. Results from this study offer new insights for research in this area.

References

- Aarons, G. A., Hurlburt, M., & Horwitz, S. M. (2011). Advancing a conceptual model of evidence-based practice implementation in public service sectors. *Administration and Policy in Mental Health and Mental Health Services Research*, 38, 4–23. doi:10.1007/s10488-010-0327-7
- Armitage, C. J., & Conner, M. (2001). Efficacy of the theory of planned behavior: A meta-analytic review. *British Journal of Social Psychology*, 40, 471–499. doi:10.1348/014466601164939
- Atkins, M. S., Frazier, S. L., Leathers, S. J., Graczyk, P. A., Talbott, E., Jakobsons, L., . . . Bell, C. C. (2008). Teacher key opinion leaders and mental health consultation in low-income urban schools. *Journal of Consulting and Clinical Psychology*, 76, 905–908. doi:10.1037/a0013036

Baker, C. N., Kupersmidt, J. B., Voegler-Lee, M. E., Arnold, D. H., & Willoughby, M. T. (2010). Predicting teacher participation in a classroom-based, integrated preventive intervention for preschoolers. *Early Childhood Research Quarterly, 25*, 270–283. doi:10.1016/j.ecresq.2009.09.005

Brouwers, A., & Tomic, W. (2000). A longitudinal study of teacher burnout and perceived self-efficacy in the classroom. *Teaching and Teacher Education, 16*, 239–253. doi:10.1016/S0742-051X(99)00057-8

Coalition for Psychology in Schools and Education. (2006, August). *Report on the Teacher Needs Survey*. Washington, DC: American Psychological Association, Center for Psychology in Schools and Education.

Codding, R. S., Feinberg, A. B., Dunn, E. K., & Pace, G. M. (2005). Effects of immediate performance feedback on implementation of behavior support plans. *Journal of Applied Behavior Analysis, 38*, 205–219. doi:10.1901/jaba.2005.98-04

Cunningham, C. E., Vaillancourt, T., Rimas, H., Deal, K., Cunningham, L., Short, K., & Chen, Y. (2009). Modeling the bullying prevention program preferences of educators: A discrete choice conjoint experiment. *Journal of Abnormal Child Psychology, 37*, 929–943. doi:10.1007/s10802-009-9324-2

Darling-Hammond, L., Chung Wei, R., Andree, A., Richardson, N., & Orphanos, S. (2009). *Professional learning in the learning profession: A status report on teacher development in the United States and abroad*. Oxford, OH: National Staff Development Council.

DeForest, P. A., & Hughes, J. N. (1992). Effect of teacher involvement and teacher self-efficacy on ratings of consultant effectiveness and intervention acceptability. *Journal of Educational and Psychological Consultation, 3*, 301–316. doi:10.1207/s1532768xjepc0304_2

Denton, C. A., & Hasbrouck, J. (2009). A description of instructional coaching and its relationship to consultation. *Journal of Educational and Psychological Consultation, 19*, 150–175. doi:10.1080/10474410802463296

Domitrovich, C. E., Bradshaw, C. P., Poduska, J. M., Hoagwood, K., Buckley, J. A., Olin, S., . . . Ialongo, N. S. (2008). Maximizing the implementation quality of evidence-based preventive interventions in schools: A conceptual framework. *Advances in School Mental Health Promotion, 1*, 6–28. doi:10.1080/1754730X.2008.9715730

Evers, W. J. G., Brouwers, A., & Tomic, W. (2002). Burnout and self-efficacy: A study on teachers' beliefs when implementing an innovative educational system in the Netherlands. *British Journal of Educational Psychology, 72*, 227–243. doi:10.1348/000709902158865

Fabiano, G. A., Pelham, W. E., Pisecco, S., Evans, S. W., Manos, M. J., Caserta, D., . . . Johnston, C. (2002, November). A nationally representative survey of classroom-based behavior modification treatment for ADHD. Poster presented at the Advancement of Behavior Therapy Conference, Reno, NV.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*, 175–191. doi:10.3758/BF03193146

Fixsen, D. L., Blase, K. A., Naom, S. F., & Wallace, F. (2009). Core implementation components. *Research on Social Work Practice, 19*, 531–540. doi:10.1177/1049731509335549

Frey, A. J., Lee, J., Small, J. W., Seeley, J. R., Walker, H. M., & Feil, E. G. (2013). The motivational interviewing navigation guide: A process for enhancing teachers' motivation to adopt and implement school-based interventions. *Advances in School Mental Health Promotion, 6*, 158–173. doi:10.1080/1754730X.2013.804334

Girio, E. L., & Owens, J. S. (2009). Teacher acceptability of evidence-based and promising treatments for children with attention-deficit/hyperactivity disorder. *School Mental Health, 1*, 16–25. doi:10.1007/s12310-008-9001-6

Greene, R., Besztercze, S. K., Katzenstein, T., Park, K., & Goring, J. (2002). Are students with ADHD more stressful to teach? Patterns of teacher stress in an elementary school sample. *Journal of Emotional and Behavioral Disorders, 10*, 79–89.

Gueldner, B., & Merrell, K. (2011). Evaluation of a social-emotional learning program in conjunction with exploratory application of performance feedback incorporating motivational interviewing technique. *Journal of Educational and Psychological Consultation, 21*, 1–27.

Hallinger, P., & Heck, R. H. (1996). Reassessing the principal's role in school effectiveness: A review of empirical research, 1980–1995. *Education Administration Quarterly, 32*, 5–44. doi:10.1177/0013161X96032001002

Hastings, R. P., & Bham, M. S. (2003). The relationship between student behavior patterns and teacher burnout. *School Psychology International, 24*, 115–127. doi:10.1177/0143034303024001905

Hudson, S. B., McMahon, K. C., & Overstreet, C. M. (2002). *The 2000 National Survey of Science and Mathematics Education: Compendium of tables*. Chapel Hill, NC: Horizon Research.

Ingersoll, R. M. (2001). Teacher turnover and teacher shortages: An organizational analysis. *American Educational Research Journal, 38*, 499–534. doi:10.3102/00028312038003499

Iwanicki, E. F., & Schwab, R. L. (1981). A cross validation study of the Maslach Burnout Inventory. *Educational and Psychological Measurement, 41*, 1167–1174.

Kam, C. M., Greenberg, M. T., & Walls, C. T. (2003). Examining the role of implementation quality in school-based prevention using the PATHS curriculum. *Prevention Science, 4*, 55–63. doi:10.1023/A:1021786811186

Kardos, S. M., Johnson, S. M., Peske, H. G., Kauffman, D., & Liu, E. (2001). Counting on colleagues: New teachers encounter the professional cultures of their schools. *Education Administration Quarterly, 37*, 250–290. doi:10.1177/00131610121969316

Kazdin, A. E. (1980). Acceptability of alternative treatments for deviant child behavior. *Journal of Applied Behavior Analysis, 13*, 259–273. doi:10.1901/jaba.1980.13-259

Kelley, M. L. (1990). *School-home notes: Promoting children's classroom success*. New York, NY: Guilford Press.

Leiter, M. P., & Durup, J. (1994). The discriminant validity of burnout and depression: A confirmatory factor analytic study. *Anxiety, Stress & Coping: An International Journal, 7*, 357–373.

Martinussen, R., Tannock, R., & Chaban, P. (2011). Teachers' reported use of instructional and behavioral management practices for students with behavior problems: Relationship to role and level of training in ADHD. *Child and Youth Care Forum, 40*, 193–210. doi:10.1007/s10566-010-9130-6

Maslach, C., & Jackson, S. E. (1986). *Maslach Burnout Inventory manual (research education)*. Palo Alto, CA: Consulting Psychologists Press.

Maslach, C., Jackson, S., & Leiter, M. P. (1996). *Maslach Burnout Inventory manual*. Palo Alto, CA: Consulting Psychologists Press.

Miller, W. R., & Rollnick, S. (2013). *Motivational interviewing, third edition: Preparing people for change*. New York, NY: Guilford Press.

Morris, S. B., & DeShon, R. P. (2002). Combining effect size estimates in meta-analysis with repeated measures and independent-groups designs. *Psychological Methods, 7*, 105–125. doi:10.1037/1082-989X.7.1.105

National Comprehensive Center for Teacher Quality. (2012). *Evaluating the effectiveness of teacher preparation programs for support and accountability: Research and policy brief*. Washington, DC: Author.

National Council on Teacher Quality. (2014). *Training our future teachers: Classroom management*. Retrieved from http://www.nctq.org/dmsView/Future_Teachers_Classroom_Management_NCTQ_Report

National Research Council. (2001). *Eager to learn: Educating our preschoolers*. Washington, DC: National Academy Press.

The New Teacher Project. (2013). *Perspectives of irreplaceable teachers*. Retrieved from http://tntp.org/assets/documents/TNTP_Perspectives_2013.pdf

Noell, G., Witt, J., Gilbertson, D., Ranier, D., & Freeland, J. (1997). Increasing teacher intervention implementation in general education settings through consultation and performance feedback. *School Psychology Quarterly, 12*, 77–88. doi:10.1037/h0088949

Owens, J. S., & Coles, E. K. (2014). *Development of strategies to increase teacher integrity in a daily report card intervention for children with or at risk for ADHD* (Institute of Education Sciences Grant R324A120272, in progress). Retrieved from <http://ies.ed.gov/funding/grantsearch/details.asp?ID=1361>

Owens, J. S., Holdaway, A. S., Zoromski, A. K., Evans, S. W., Himawan, L. K., Girio-Herrera, E., & Murphy, C. E. (2012). Incremental benefits of a daily report card intervention over time for youth with disruptive behavior. *Behavior Therapy, 43*, 848–861. doi:10.1016/j.beth.2012.02.002

Pelham, W. E., & Fabiano, G. A. (2008). Evidence-based psychosocial treatment for attention-deficit/hyperactivity disorder. *Journal of Clinical Child & Adolescent Psychology, 37*, 184–214. doi:10.1080/15374410701818681

Reinke, W. M., Frey, A. J., Herman, K., & Thompson, C. (2014). Improving engagement and implementation of interventions for children with behavior problems in home and school settings. In H. Walker & F. Gresham (Eds.), *Handbook of evidence-based practices for students having emotional and behavioral disorders* (pp. 432–445). New York, NY: Guilford Press.

Reinke, W. M., Herman, K. C., Darney, D., Pitchford, J., Becker, K., Domitrovich, C., & Ialongo, N. (2012). Using the classroom check-up model to support implementation of PATHS to PAX. *Advances in School Mental Health Promotion, 5*, 220–232. doi:10.1080/1754730X.2012.707441

Reinke, W. M., Herman, K. C., & Sprick, R. (2011). *Motivational interviewing for effective classroom management: The classroom check-up*. New York, NY: Guilford Press.

Rimm-Kaufman, S. E., & Sawyer, B. E. (2004). Primary-grade teachers' self-efficacy beliefs, attitudes toward teaching, and disciplines and teaching practice priorities in relation to the "responsive classroom" approach. *The Elementary School Journal, 104*, 321–341. doi:10.1086/499756

Rogers, E. M. (2003). *The diffusion of innovations* (5th ed.). London, England: Free Press.

Shernoff, E. S., Martinez-Lora, A. M., Frazier, S. L., Jakobsons, L. J., & Atkins, M. S. (2011). Teachers supporting teachers in urban schools: What iterative research designs can teach us. *School Psychology Review, 40*, 465–485.

Smith, T. M., & Ingersoll, R. M. (2004). What are the effects of induction and mentoring on beginning teacher turnover? *American Educational Research Journal, 41*, 681–714. doi:10.3102/00028312041003681

Solomon, B. G., Klein, S. A., & Politylo, B. C. (2012). The effect of performance feedback on teachers' treatment integrity: A meta-analysis of the single-case literature. *School Psychology Review, 41*, 160–175.

Tschannen-Moran, M., & Hoy, W. (2001). Teacher efficacy: Capturing an elusive construct. *Teaching and Teacher Education, 17*, 783–805. doi:10.1016/S0742-051X(01)00036-1

U.S. Department of Education. (2009). *Common core of public school data*. Retrieved from <http://nces.ed.gov/ccd/schoolsearch/>

Vannest, K. J., Davis, J. L., Davis, C. R., Mason, B. A., & Burke, M. D. (2010). Effective intervention for behavior with a daily behavior report card: A meta-analysis. *School Psychology Review, 39*, 654–672.

Worley, J. A., Vassar, M., Wheeler, D. L., & Barnes, L. L. (2008). Factor structure of scores from the Maslach Burnout Inventory: A review and meta-analysis of 45 exploratory and confirmatory factor-analytic studies. *Educational and Psychological Measurement, 68*, 797–823. doi:10.1177/0013164408315268

Yoon, K. S., Duncan, T., Lee, S. W.-Y., Scarloss, B., & Shapley, K. (2007). *Reviewing the evidence on how teacher professional development affects student achievement* (Issues & Answers Report, REL 2007-No. 033). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Southwest. Retrieved from <http://ies.ed.gov/ncee/edlabs>

(Appendix follows)

Appendix

ISQ Construction and Validation Information

Validation of the Intervention Support Questionnaire (ISQ)

To ensure that vignette descriptions accurately described intended constructs and procedures, vignette validation was conducted in a two-step procedure. After we constructed preliminary descriptions of the child with ADHD, the DRC intervention, and the four TCCs, feedback was obtained from graduate students and faculty members familiar with classroom interventions and teacher consultation for students with disruptive behaviors. This feedback guided the first iteration of revisions. During the second step of vignette validation, revised vignettes were distributed for review by established researchers who specialize in development and evaluation of school-based treatments for children with disruptive behaviors ($n = 3$) and educators currently employed in a local school district ($n = 4$). Research experts had over 60 combined years of research experience on school-based treatments for children with disruptive behaviors, had published over 200 peer-reviewed journal publications and have received over 20 federal grants for researching school-based interventions for students with behavior problems. Local educators included a school mental-health professional with over 20 years of experience and three classroom teachers with over 30 years combined experience. All participants in the second step of validation were asked to rate: (1) how severe the child's behavior was on a scale from 0 (*no problem*) to 6 (*extreme problem*), (2) how disruptive the child's behavior would be to a classroom environment on a scale from 0 (*no problem*) to 6 (*extreme problem*), and (3) the extent to which the intervention was appropriate for the child's behavioral difficulties on a scale from 0 (*strongly disagree that this is an appropriate intervention*) to 6 (*strongly agree this is an appropriate intervention*). Ratings indicated that the vignettes validly represent the goals described above. Namely, ratings of child severity ($M = 4.28$, $SD = .48$) and disruptiveness ($M = 4.42$, $SD = .53$) fell in the intended range, and ratings of the intervention indicated that respondents found it to be an appropriate intervention for the child's difficulties ($M = 5.29$, $SD = .76$).

Additionally, to ensure differentiation between TCC descriptions, researchers, but not educators, were asked to rate the extent to which each TCC's description represented their professional conceptualizations of each of the TCCs on a scale from 0 (*not at all*) to 4 (*to a great extent*). A dummy condition label ("Consultation with Stress Reduction Techniques") was also included so that raters could not simply "match" the descriptions to a label when making their ratings. These raters reported that the descriptions largely matched their professional conceptualizations of the TCC labels, providing further validation for the ISQ.

tion for the ISQ. More specifically, 3 of 4 TCC descriptions received ratings of 4 for the intended label, and one received a mean rating of 3.33, indicating that our descriptions represented the intended TCC label to a great extent. Further, no TCC description was rated greater than a 1.0 for any other TCC label, indicating that the TCC conditions were well differentiated.

Intervention Support Questionnaire

Observation and Performance Feedback Description

To support you in using the new intervention, the counselor agrees to observe your classroom two times per month and have meetings with you to discuss how the intervention is going. In these meetings she will help you figure out how to best deliver all the components of the intervention and troubleshoot problems specific to Sam and his response to the intervention. More specifically, she will review and graph the data on how Sam is progressing, discuss your strengths in the use of the intervention, discuss any ideas for improving intervention use and outcomes, and answer any questions or concerns you may have. This would be available to you throughout the school year.

Motivational Interviewing Description

The counselor talks with you to obtain information about Sam's behaviors and discuss the new intervention. In the process, you share concerns you have about implementing the intervention. You feel that your current schedule is full, and you have too many kids in the classroom to be able to provide special attention to any one. The counselor facilitates an interview and discussion that helps you explore your hesitations. During this discussion you explain your hesitations (e.g. "I'd like to help Sam, but I am concerned about the time this intervention will take"; "I'm not sure I can handle one more thing on top of all the things I already do"). You then discuss the potential positives of the program: the classroom could be calmer, you could get back some of the instruction time you currently use to redirect Sam, and he may be less prone to arguing with you. Through this process you come to identify possible benefits of implementing the intervention and the school counselor agrees to be available to answer any more questions you may have or help generate ideas about how to deal with potential obstacles.

Key Opinion Leader Description

Before reading the vignette below, please think of a colleague whom you respect and whom you typically turn to for advice and/or guidance regarding disruptive student behavior. Insert this person's name in the blank spaces as you read.

The counselor informs you that _____ tested this intervention in his/her classroom last year and will meet with you to discuss the intervention. Because you respect this colleague's opinion and experience, you then meet with _____ and discuss the program. _____ specifically points out strategies and techniques that worked particularly well for him/her in the past year and says the program helped a child in his/her class. If you decide to use this program, _____ will be available informally as needed throughout the school year (e.g. before or after school, during planning periods) to discuss progress and help you to use the intervention.

Training-as-Usual Description

The mental health professional says that there is a new program which may be helpful for Sam. You are encouraged to attend the workshop to receive training. The workshop will be primarily PowerPoint based and taught through a two hour session. The session will include a lecture, a role play and a question and answer session at the end. You are given a manual that has information about the techniques, quotes from teachers about the

program and some tips for putting it in use. You are also given a website which has information, as well as worksheet examples available for download.

1. Now that you have read all the descriptions of consultation and support strategies, please rank the strategies in order of your preference. Place a "1" next to the support strategy that would make it *most likely* that you would use the intervention. Place a "2" next to the support description that would be second most likely, and so on. Please rank all four strategies. Assume that you have to use support strategies, even if your answer on previous questions was that you would never use the intervention for any strategy.

Rank

- _____ Support Strategy A
- _____ Support Strategy B
- _____ Support Strategy C
- _____ Support Strategy D

Received April 22, 2013
Revision received May 1, 2014
Accepted May 26, 2014 ■

Earlier School Start Times as a Risk Factor for Poor School Performance: An Examination of Public Elementary Schools in the Commonwealth of Kentucky

Peggy S. Keller, Olivia A. Smith, Lauren R. Gilbert,
Shuang Bi, and Eric A. Haak
University of Kentucky

Joseph A. Buckhalt
Auburn University

Adequate sleep is essential for child learning. However, school systems may inadvertently be promoting sleep deprivation through early school start times. The current study examines the potential implications of early school start times for standardized test scores in public elementary schools in Kentucky. Associations between early school start time and poorer school performance were observed primarily for schools serving few students who qualify for free or reduced-cost lunches. Associations were controlled for teacher-student ratio, racial composition, and whether the school was in the Appalachian region. Findings support the growing body of research showing that early school start times may influence student learning but offer some of the first evidence that this influence may occur for elementary school children and depend on school characteristics.

Keywords: sleep, start time, school performance, free lunch

Adequate high-quality sleep is important for the daytime functioning of children (Paavonen et al., 2000). Consequences of inadequate sleep include irritability, emotional dysregulation, impulsivity, difficulties with attention, and poorer cognitive performance (Curcio, Ferrara, & De Gennaro, 2006). It is therefore important to understand factors that may hinder child sleep. For children, wake times are partially determined by school start times; to attend school, children must wake early enough to get ready and be transported to the school (Wolfson, Spaulding, Dandrow, & Baroni, 2007). By curtailing the sleep period, earlier school start times may reduce the amount of sleep children can obtain (Dexter, Bijwadia, Schilling, & Applebaugh, 2003) and lead to sleep deprivation. Thus, early school start times may indirectly lead to poor school performance by causing sleep deprivation (Dworak, Schieler, Bruns, & Struder, 2007). However, a large scale investigation of the potential impact of public school start times on academic achievement is lacking, and very little research has examined the impact of start times for elementary school students. The purpose of the current study is to address these gaps by examining associations between public elementary school start times and school performance measures in the public schools of Kentucky.

Sleep problems have been linked to poor school performance and low attendance rates (Sadeh, Gruber, & Raviv, 2003). For example, sleep quality and quantity in school children are related to declarative and procedural learning (Curcio et al., 2006). Daytime sleepiness is associated with executive functioning problems such as poor concentration and difficulty focusing attention (Anderson, Storfer-Isser, Taylor, Rosen, & Redline, 2009; Buckhalt, El-Sheikh, Keller, & Kelly, 2009; El-Sheikh, Buckhalt, Keller, Cummings, & Acebo, 2007). Shorter sleep duration is also linked to working memory capacity and memory consolidation (Kopasz et al., 2010), cognitive abilities that are very important for academic performance. A recent meta-analysis of over a century of research demonstrated a small but reliable association between children's longer sleep duration and better performance on cognitive tasks and higher academic achievement (Astill, Van der Heijden, Van IJzendoorn, & Van Someren, 2012). Another recent meta-analysis suggests that sleepiness and sleep duration are related to child school performance (Dewald, Meijer, Oort, Kerkhof, & Bogels, 2010). Further, treatment of child sleep disorders is associated with improvements in attention (Chervin et al., 2006).

Early school start times are a potential cause of child and adolescent sleep deprivation because they curtail the sleep period (Knutson & Lauderdale, 2009). There are now a number of studies documenting the link between early school start times and lower sleep amount and daytime sleepiness in adolescents (e.g., Dexter et al., 2003; Epstein, Chillag, & Lavie, 1998; Li et al., 2013; Wahlstrom, 2002). For example, a change in high school start times from 8:25 a.m. to 7:20 a.m. was associated with student sleep deprivation and greater daytime sleepiness (Carskadon, Wolfson, Acebo, Tzischinsky, & Seifer, 1998). Wolfson et al. (2007) examined two middle schools, one starting classes at 7:15 a.m. (School E) and one starting at 8:37 a.m. (School L). Adolescents attending School E had significantly more daytime sleepiness and reported

This article was published Online First June 16, 2014.

Peggy S. Keller, Olivia A. Smith, Lauren R. Gilbert, Shuang Bi, and Eric A. Haak, Department of Psychology, University of Kentucky; Joseph A. Buckhalt, Department of Counselor Education, Counseling Psychology, and School Psychology, Auburn University.

This research was supported by the Chellgren Center and the Office of Undergraduate Research of the University of Kentucky.

Correspondence concerning this article should be addressed to Peggy S. Keller, Department of Psychology, University of Kentucky, Lexington KY 40506. E-mail: peggy.keller@uky.edu

37 fewer minutes of total sleep than adolescents attending School L. Further, adolescents attending School E had 4 times more tardies than those attending School L. Owens, Belon, and Moss (2010) examined a 30-min delay in start time at a private high school and observed a 45-min increase in average sleep duration, reduced percentage of sleep deprived students, and declines in daytime sleepiness.

Sleep deficits associated with early school start times may translate into poor school performance. A 1-hr delay in middle school start times (8:30) was associated with improved student performance on tests of attention and impulsivity compared to students attending school at the regular time (7:30); these improvements disappeared after the experimental group returned to the normal start time (Lufi, Tzischinsky, & Hadar, 2011). When schools in Wake County, North Carolina, delayed school start times, Edwards (2012) compared student performance on standardized tests of math and reading before (1999) and after (2006) the delay. A 1-hr delay in middle schools and high schools was related to improved test scores on math and reading (Edwards, 2012). Effects were especially strong for students with lower test scores. Notably, this study found no effects of school start times on elementary school students' performance.

Despite the strengths of these prior research studies, there are some notable gaps in research on school start times and academic performance. First, the majority of prior studies have been case studies or studies of schools in only one school district (although see Li et al., 2013, for an exception). This makes it difficult to judge the widespread impact of school start times on academic performance. It also leads to the second gap in research: There is currently little understanding of how school start times relate to student performance in schools with differing characteristics. Few studies have examined moderators of the association between school start times and child or adolescent functioning, and none have examined socioeconomic status variables as moderators. Finally, research has almost exclusively considered middle and high school students. School start times are proposed to be more influential for adolescents because of biological changes in sleep-wake regulation associated with puberty (Crowley, Acebo, & Carskadon, 2007). On the basis of evidence that early school start times are harmful for adolescents, some school districts have chosen to push middle and high school start times later and make elementary school start times earlier to retain staggered busing strategies (Kirby, Maggi, & D'Angiulli, 2011). It is therefore critical to investigate the impact of early school start times on elementary school students.

The current study addresses these research gaps. We examine associations between school start times and average standardized test scores for elementary schools in all public school districts in Kentucky. We chose not to include middle and high schools in our analysis because we found very little variability in middle and high school start times in Kentucky. We hypothesize that schools with earlier start times will have lower average student test scores and poorer school performance. We also examine two school differences as moderators of the association between school start time and student test scores: county designation as Appalachian and the percentage of students receiving free or reduced-cost lunches.

The Appalachian region includes the vast majority of eastern Kentucky. Appalachian counties are known for their low economic status, including high poverty levels and very few job opportuni-

ties (de Young, 1985). Although the Appalachian region has been improving in terms of academic performance and employment rates, it still lags behind non-Appalachian areas (Shaw, De Young & Rademacher, 2004; Wilson & Gore, 2009). For example, Appalachian counties have high school dropout rates that are double the national average (Laird, Cataldi, KewalRamani, & Chapman, 2008), making them the lowest completion rates in the United States (Ziliak, 2012). Because Appalachian schools experience greater problems, they may be especially susceptible to the possible effects of early school start times. We therefore hypothesize that associations between school start times and student test scores will be stronger for Appalachian school districts.

School start times may also have an important impact in schools serving economically disadvantaged populations. There is a well-documented achievement gap between poor and middle class students, and this gap has been steadily increasing over the last 70 years (H. F. Ladd, 2012). There are likely numerous reasons for this gap, including poorer student health, less access to high quality preschools, residential mobility or lack of mobility (e.g., it may be difficult for poor parents to move into areas with high quality schools), and the inability to afford expensive extracurricular activities that enhance cognitive development (Evans, 2004). Sleep may therefore be especially important for economically disadvantaged students (Buckhalt, 2011). A common indicator of poverty is eligibility for free or reduced-cost school lunch. We hypothesize that the association between school start times and test scores will be stronger for those schools with a higher percentage of students receiving free or reduced-cost lunches.

Method

Data were collected for all eligible public elementary schools in Kentucky. Schools were considered ineligible if they were vocational schools, alternative schools, schools that only included prekindergarten through the second grade (test data are not available for these grades), private schools, special education schools, and schools in juvenile justice centers. Two elementary schools were removed from analyses because their start time was 1:40 p.m. We were unable to determine the start time for one elementary school. The resulting sample included 718 elementary schools.

School start time data were collected via school websites or by calling the school office. Other variables were obtained via the Kentucky Department of Education website (<http://education.ky.gov>). Variables included in the study are listed below. Data are from the 2011–2012 school year (Kentucky Department of Education, 2011, 2012). Means and standard deviations are provided in Table 1.

School start times. Start times were computed as minutes since midnight.

Novice, Apprentice, Proficient, Distinguished (NAPD) scores. Each school had scores evaluating student performance on the Kentucky Performance Rating for Educational Progress (K-PREP) assessment in each of the following domains: reading, mathematics, science, social studies, and writing. These scores are referred to as NAPD scores because they were based on the percentages of children classified as novice, apprentice, proficient, and distinguished, based on cutoff scores (see <http://www.education.ky.gov> for details). K-PREP exams were administered in third and fourth grades. The possible range of the K-PREP scores was 0–30 for

Table 1
Means, Standard Deviations, and Other Descriptive Statistics of Study Variables

Variable	Elementary <i>M</i> (<i>SD</i>)	Middle <i>M</i> (<i>SD</i>)	High <i>M</i> (<i>SD</i>)
Start time	8:05 AM (35 min)	8:00 AM (20 min)	8:01 AM (18 min)
Minimum	7:00 AM	7:20 AM	7:20 AM
Maximum	9:10 AM	9:05 AM	9:05 AM
Schools starting at:			
7:00–7:19	1 (0.1%)	0	0
7:20–7:59	350 (48.7%)	151 (45.3%)	90 (39.0%)
8:00–8:29	224 (31.2%)	150 (45.1%)	121 (52.4%)
8:30–8:59	41 (5.7%)	22 (6.6%)	17 (7.4%)
9:00–9:10	102 (14.2%)	10 (3%)	3 (1.2%)
NAPD Language	66.24 (17.85)	30.77 (54.02)	66.06 (26.42)
NAPD Reading	62.01 (13.36)	58.94 (14.65)	55.46 (20.01)
NAPD Math	60.45 (13.40)	58.67 (15.41)	48.40 (35.19)
NAPD Writing	56.78 (12.46)	63.52 (16.90)	63.41 (18.54)
NAPD Science	88.58 (13.21)	74.69 (33.26)	46.99 (34.30)
NAPD Social Studies	78.47 (14.96)	72.85 (33.48)	44.73 (32.51)
Attendance rate	95.20 (1.23)	94.06 (10.68)	93.27 (1.83)
Retention rate	0.437 (.949)	0.231 (7.79)	3.33 (3.29)
Graduation rate			71.03 (37.45)
College transition rate			53.28 (25.56)
Student-teacher ratio	15.30 (2.11)	15.09 (9.18)	14.91 (10.90)

each subject at each grade, but cutoff scores differed by subject and grade. Table 2 presents details regarding cutoffs for classifications and grades in which the tests were administered. NAPD scores were computed as follows: Schools received 1 point for every percentage point of students scoring proficient or distinguished (for a maximum score of 100); half a point was awarded for each percentage point of students scoring apprentice. NAPD scores are therefore continuous, and higher scores represent better school performance.

School rank. This variable is the percentile rank of a school based on overall school performance, ranging from 0 to 100. Higher percentile rank indicates better school performance. Schools are ranked against other schools of their level (e.g., other elementary schools).

Attendance rate. Schools provided the percentage of enrolled students in attendance for every school day to the Kentucky Department of Education. The attendance rate is the average attendance percentage across the entire school year.

Retention rate. The retention rate is the percentage of a school's students who have been required to repeat a grade.

Appalachian county (APPALACHIAN). This variable identifies whether the school is located in a county that has been designated as Appalachian according to the Appalachian Regional Commission (<http://www.arc.gov/about/index.asp>). Fifty-four of the 120 counties in Kentucky are designated as Appalachian.

Free and reduced-cost lunches (FREELUNCH). This is the percentage of students in the school receiving free or reduced-cost lunches.

Teacher-student ratio (TSRATIO). The variable reflects the average number of students per teacher.

Percentage African American (AFRICAN AMERICAN). The percentage of students who are African American in a given school is reflected in this variable. The average percentage across all elementary schools was 9.14% (*SD* = 14.56%) and ranged from 0.0% to 76.0%. However, 65% of schools were 5% or less African American. Only 2.9% of schools served a

population of students in which the majority was African American.

Percentage Hispanic (HISPANIC). The percentage of students who are Hispanic in a given school is reflected in this variable. The average percentage across all elementary schools was 4.70% (*SD* = 6.68%). However, 71.3% of schools were 5% or less Hispanic. Only two schools (< 1%) served a population of students in which the majority was Hispanic.

Data Analyses

Because schools were nested within county (in Kentucky, there is one school district for each county), schools within the same county were not independent of each other and multilevel modeling was required for data analysis (see Raudenbush & Bryk, 2002 for a detailed overview of this statistical procedure). Multilevel modeling for nested data and similar procedures are common in educational research (e.g., Dettmers, Trautwein, Lüdtke, Kunter, & Baumert, 2010; Goddard & Goddard, 2001; Shen, Leslie, Spybrook, & Ma, 2012; Wenglinsky, 2002), including research on school start times (Edwards, 2012). In multilevel modeling, within-county variability is partitioned from between-county variability. At Level 1, the within-county level, dependent variables (e.g., NAPD scores) for schools (I) in counties (J) are modeled as a function of an intercept (B_{j0} ; the expected value of the dependent variable when there are scores of zero on the independent variables included in the Level 1 model) and the effects of independent variables that vary from school to school within the same county (e.g., school start times; B_{j1}):

$$\begin{aligned}
 \text{NAPDMATH}_{IJ} = & B_{j0} + B_{j1} (\text{STARTTIME}_I) \\
 & + B_{j2} (\text{FREELUNCH}_I) + B_{j3} (\text{TIMEXLUNCH}_I) \\
 & + B_{j4} (\text{AFRICAN AMERICAN}_I) \\
 & + B_{j5} (\text{HISPANIC}_I) + B_{j6} (\text{TSRATIO}_I).
 \end{aligned}$$

Table 2
Per Grade Administration of Standardized Tests and Total Score Ranges Per Student Classification

Subject	Novice	Apprentice	Proficient	Distinguished
Grade 3				
Reading	0-8	9-16	17-23	24-30
Mathematics	0-9	10-16	17-24	25-30
Grade 4				
Reading	0-8	9-16	17-23	24-30
Mathematics	0-8	9-16	17-23	24-30
Science	0-9	10-17	18-24	25-30
Language Mechanics	0-9	10-17	18-25	26-30
Grade 5				
Reading	0-9	10-17	18-24	25-30
Mathematics	0-8	9-16	17-25	26-30
Social Studies	0-10	11-18	19-25	26-30
Writing	0-9	10-17	18-24	25-30
Grade 6				
Reading	0-9	10-17	18-24	25-30
Mathematics	0-8	9-15	16-25	26-30
Writing	0-9	10-17	18-24	25-30
Language Mechanics	0-9	10-17	18-24	25-30
Grade 7				
Reading	0-9	10-14	15-20	21-30
Mathematics	0-8	9-14	15-22	23-30
Science	0-9	10-16	17-21	22-30
Grade 8				
Reading	0-8	9-15	16-21	22-30
Mathematics	0-9	10-15	16-22	23-30
Social Studies	0-9	10-17	18-25	26-30
Writing	0-10	11-18	19-25	26-30
Grade 9				
Reading	0-9	10-17	18-24	25-30
Grade 10				
Mathematics	0-9	10-15	16-22	23-30
Writing	0-9	10-17	18-24	25-30
Language Mechanics	0-9	10-17	18-24	25-30
Grade 11				
Writing	0-8	9-16	17-24	25-30
Science	0-9	10-15	16-22	23-30
Grade 12				
Social Studies	0-9	10-18	19-25	26-30

The above equation illustrates that we examined associations between start times and school performance, controlling for teacher-student ratio, percentage of students identified as African American, and percentage of students identified as Hispanic. Coefficients for the independent variables are interpreted in essentially the same way as regression coefficients. Interactions between Level 1 variables can be entered (B_{j3}) and indicate whether level one coefficients vary based on the values of other Level 1 variables.

In essence, each county has its own regression equation. At Level 2, the between-county level, each of the coefficients at Level 1 is

modeled as a linear function of an intercept (e.g., π_{10} ; the expected value of the Level 1 coefficient for schools with values of zero on the other variables entered into the Level 2 equation) and the effects of independent variables that only vary from county to county and not within county (e.g., π_{20} , designation of Appalachian county):

$$B_{j0} = \pi_{10} + \pi_{20} (\text{APPALACHIAN}_j)$$

$$B_{j1} = \pi_{11} + \pi_{21} (\text{APPALACHIAN}_j)$$

$$B_{j2} = \pi_{12}$$

$$B_{j3} = \pi_{13}$$

$$B_{j4} = \pi_{14}$$

$$B_{j5} = \pi_{15}$$

$$B_{j6} = \pi_{16}.$$

Coefficients for the Level 2 predictors in the top equation can be interpreted as the first-order effects of the Level 2 variables on the dependent variable. That is, the coefficient π_{20} in the top equation above represents the effect of Appalachian county designation on NAPD math scores. Coefficients for the Level 2 predictors of the other Level 1 coefficients can be interpreted as moderation effects: They provide information concerning whether the Level 1 coefficient varies based on between-county variables. That is, the coefficient π_{21} indicates whether the effect of school start times on NAPD math scores depends on whether the school is located in an Appalachian county. Level 2 independent variables could be added to any of the Level 2 models, but such effects were not of interest in the current study. The coefficients π_{12} through π_{16} therefore indicate the average effects across all counties of the percentage of students receiving free or reduced-cost lunches, the interaction between this variable and school start times, AFRICAN AMERICAN, HISPANIC, and TSRATIO, respectively. Estimates of coefficients and their standard errors are only provided at Level 2. Only unstandardized coefficients are presented.

Separate models were fit predicting each NAPD subject score, school rank, attendance rate, and retention rate. School rank is an ordinal variable. However, alternative modeling techniques for estimating nested ordinal variables is beneficial primarily when there are seven or fewer categories (Bauer & Sterba, 2011). School rank had 99 different categories. We therefore use traditional multilevel modeling for these data. All continuous independent variables were mean centered before computing cross products. Designation of county as Appalachian (APPALACHIAN) was a dummy variable coded as 0 for non-Appalachian and 1 for Appalachian. Separate models were also fit for interactions between school start times and either FREELUNCH or APPALACHIAN. Effects were considered significant if $p < .05$. Significant interactions were plotted at $\pm 1 SD$ from the mean for school start times and FREELUNCH or for Appalachian/non-Appalachian counties. Significant interactions were probed using online utilities available at <http://www.quantpsy.org> (Preacher, Curran, & Bauer, 2006).

Results

Interactions Between School Start Times and FREELUNCH

Several significant interactions between elementary school start times and FREELUNCH were observed (see Table 3). The

Table 3

Model Results for Interactions Between Elementary School Start Times and Fraction of Students Receiving Free or Reduced-Cost Lunches

Variable	NAPD							Attendance rate	Retention rate
	Language	Reading	Math	Science	Social Studies	Writing	School rank		
Intercept									
Intercept (π_{10})	68.145***	62.875***	62.481***	90.430***	80.10***	57.719***	52.937***	95.718***	0.365***
APPALACHIAN (π_{10})	-9.126***	-6.863***	-6.354***	-8.814***	-6.288**	-4.963**	-16.165***	-1.520***	0.313**
TSRATIO									
Intercept	1.520***	1.103***	.673**	.851***	1.226***	.798**	1.777***	.080***	-.041*
AFRICAN AMERICAN									
Intercept	-.523***	-.472***	-.417***	-.432***	-.413***	-.324***	-1.031***	-.005*	-.001
HISPANIC									
Intercept	-.487***	-.495***	-.402***	-.410***	-.347***	-.162*	-.692***	-.011**	-.009**
School Start Time									
Intercept (π_{11})	.059*	.038	.044*	.017	.058**	.055**	.137**	.002	.002*
FREE LUNCH									
Intercept (π_{12})	-.637*	-.705***	-.562**	.001	-.248	-.301	-.602	-.009	-.015
Start Time \times LUNCH									
Intercept (π_{13})	-.017*	-.015***	-.012*	-.010*	-.010**	-.013**	-.029***	-.001*	.000

Note. Columns indicate the dependent variable being predicted. Statistical notation provided in parentheses corresponds to the equations provided in the analysis section.

* $p < .05$. ** $p < .01$. *** $p < .001$.

interaction predicted NAPD Language scores, $\pi_{13} = -.017$, $p < .05$; NAPD Reading scores, $\pi_{13} = -.015$, $p < .001$; NAPD Science scores, $\pi_{13} = -.010$, $p < .05$; NAPD Math scores, $\pi_{13} = -.012$, $p < .05$; NAPD Social Studies scores, $\pi_{13} = -.010$, $p < .01$; NAPD Writing scores, $\pi_{13} = -.013$, $p < .01$; school rank, $\pi_{13} = -.029$, $p < .001$; and school attendance rate, $\pi_{13} = -.001$, $p < .05$.

Interactions were plotted and were all nearly identical (see Figures 1 and 2 for examples). Results of probing the interactions are also shown in Table 4. The first two rows show the simple slopes for the effect of school start time on the dependent variable (see column heading) for lower and higher values of FREELUNCH. The bottom two rows illustrate the expected difference in the dependent variable for schools starting 1 hr later than another school. In all cases, there was a significant association between school start times and school performance only for schools with a lower percentage of students receiving free or reduced-cost lunches (e.g., school with more middle and upper class students). The difference in NAPD scores associated with a 1-hr difference in school start time ranged from 3 to almost 7 points. A 1-hr difference in school start time was associated with school rank improved by 14 percentile points, and an attendance rate that was .32 units higher.

Interactions Between School Start Times and APPALACHIAN

No significant interactions were observed.

Main Effects of School Start Times

Only one main effect of school start times that was not qualified by an interaction was observed. Later school start times were associated with higher retention rates, $\pi_{11} = .002$, $p < .01$. Every additional minute later in the school start time increased retention

rates by 0.2%. A 1-hr difference in school start time would therefore be related to a 12% difference in retention rate.

Discussion

Prior research has indicated an association between early school start times and less total sleep time, more daytime fatigue and

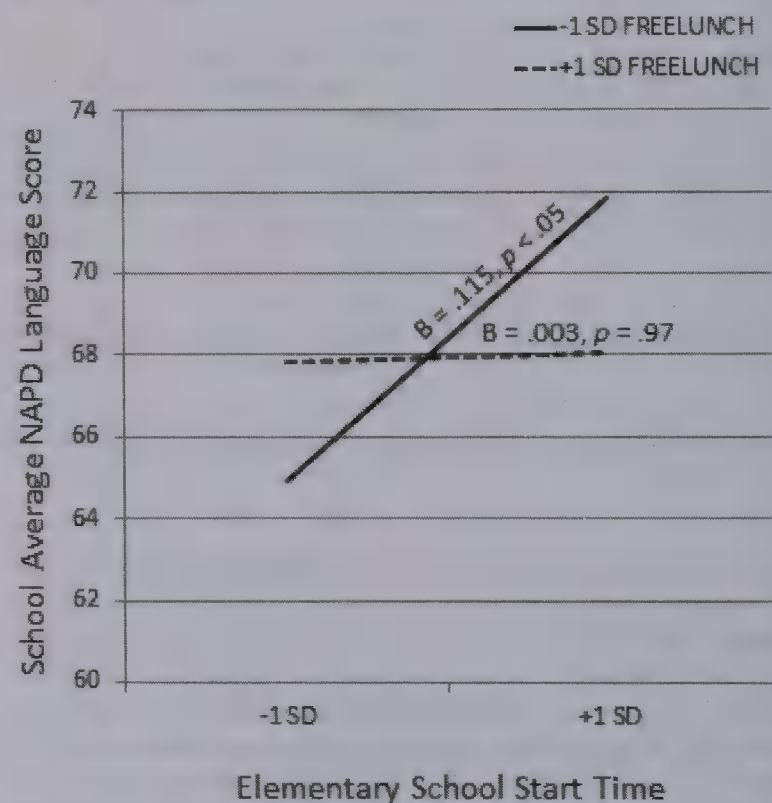


Figure 1. Interaction between Elementary School Start Time and FREELUNCH.

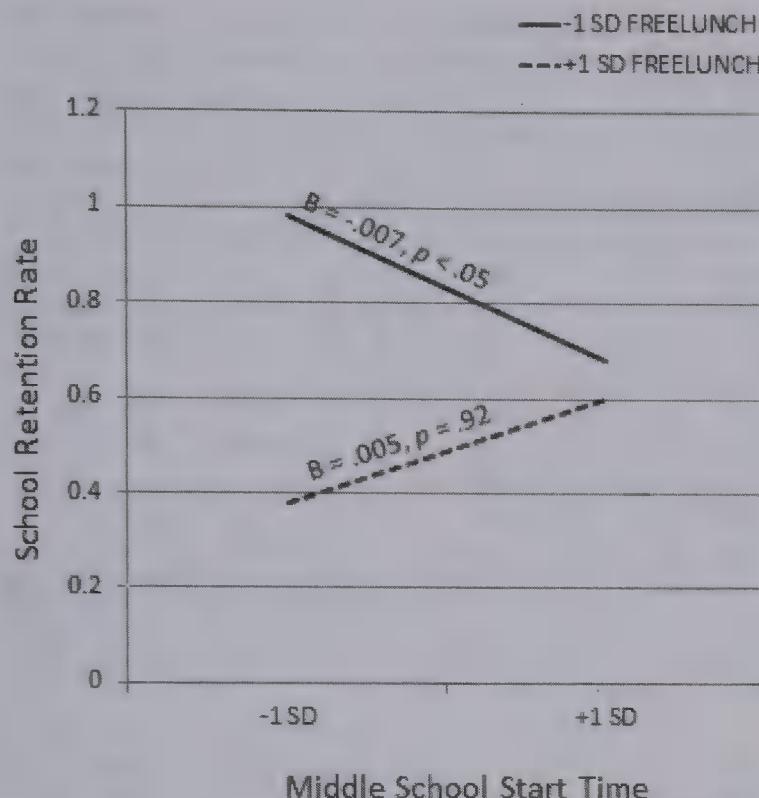


Figure 2. Interaction between Middle School Start Time and FREE LUNCH.

sleepiness, more school tardiness, and lower school academic performance (Epstein et al., 1998; Owens et al., 2010; Wahlstrom, 2002; Wolfson et al., 2007). However, no study to our knowledge has studied these associations between school start time, attendance rates, and academic performance on a statewide level. The present study investigated relations between school start times and a number of school performance standards in public elementary schools in Kentucky. We had two main hypotheses: (a) Earlier school start times will be associated with lower standardized test scores, poorer attendance, higher retention rates, lower school rank, and school underperformance; and (b) earlier start times will be especially risky for school performance standards in more disadvantaged schools, including Appalachian schools and schools with a higher percentage of students receiving free or reduced-cost lunches. Unexpectedly, findings indicated the earlier school start times were related to lower school performance predominantly for

elementary schools with fewer students receiving free or reduced-cost lunches. No differences in associations between Appalachian and non-Appalachian counties were observed.

For those schools for which an association was found, earlier start times were related to poorer test scores, lower school rank, and more student absences. These findings are consistent with previous research (Epstein et al., 1998; Wahlstrom, 2002; Wolfson et al., 2007). The relationship between earlier start times and poorer academic performance may be explained by the physical, behavioral, and psychological ramifications of sleep deprivation. Earlier start times may lead to student sleep deprivation by placing constraints on the amount of sleep a child or adolescent is able to obtain (Dexter et al., 2003; Epstein et al., 1998; Wolfson & Carskadon, 1998; Wolfson et al., 2007). Students may therefore lose the ability to remain alert and focused in the classroom (Durmer & Dinges, 2005; Epstein et al., 1998). Sleep deprivation increases hyperactivity and behavioral dysregulation, impairing students' academic functioning (Dworak et al., 2007; Beebe, 2011; Wolfson & Carskadon, 1998). Sleep problems are also associated with asthma (Kakkar & Berry, 2009), compromised cardiovascular health (Cappuccio, Cooper, D'Elia, Strazzullo, & Miller, 2011), gastrointestinal problems (Chen, Liu, Yi, & Orr, 2011), and reduced effectiveness of the immune system (Bryant, Trinder, & Curtis, 2004; Irwin et al., 1996). Therefore, sleep deprivation resulting from early school start times may increase the frequency, severity, and duration of illness, resulting in increased rates of absenteeism.

Findings clearly show that—at least for middle and upper class students—earlier school start times can be associated with poorer school performance in elementary schools. The implication is that research on school start times should not focus exclusively on adolescents. Sufficient sleep is of critical importance across development (Fallone, Owens, & Deane, 2002). According to the National Sleep Foundation 2004 Sleep in America Poll, more than 25% of school-age children (first grade to fifth grade) obtain less than the recommended daily amount of sleep. Modern-day elementary school children may be taking on additional responsibilities, extracurricular activities, and/or entertainment opportunities that delay regular weeknight bedtimes. The use of media by children (e.g., television, video games) has been identified as especially problematic for delaying bedtimes, increasing sleep onset latency, and decreasing the amount of total sleep time

Table 4

Results of Probing Interactions Between School Start Times and Percentage of Students Receiving Free or Reduced-Cost Lunches

Effects and differences of start times	NAPD							Attendance rate
	Language	Reading	Math	Science	Social Studies	Writing	School rank	
Estimated effect of school start times								
Schools with lower FREE LUNCH	.115*	.088*	.050*	.084*	.091*	.098**	.233***	.002*
Schools with higher FREE LUNCH	.003	-.012	-.016	.004	.025	.012	.041	-.001
Difference in schools starting 1 hr apart								
Schools with lower FREE LUNCH	6.90	6.23	3.01	5.03	5.48	5.90	14.01	0.32
Schools with higher FREE LUNCH	0.18	-.72	-.96	0.24	1.50	0.72	2.46	-.06

Note. The first two rows show the simple slopes for the effect of school start time on the dependent variable (see column heading) for lower and higher values of the moderator (FREE LUNCH). The bottom two rows illustrate the expected difference in the dependent variable for schools starting 1 hr later than another school.

* $p < .05$. ** $p < .01$. *** $p < .001$.

obtained (National Sleep Foundation, 2011; Owens et al., 1999). As a result, early school start times may affect student performance even before the puberty-related delay in sleep phase.

Of particular concern is that the growing public support for delaying middle and high school start times is often at the expense of making elementary school start times earlier. Indeed, this has already occurred in two counties in Kentucky (Fayette and Jessamine; National Sleep Foundation, 2005a, 2005b). This is often done in order to preserve staggered bus scheduling (Kirby et al., 2011). Our findings suggest that these policy changes may simply be shifting the problem from adolescents to younger children, instead of eliminating it altogether. On the one hand, elementary school children are not experiencing the puberty-related phase shift in sleep-wake regulation. Therefore, earlier bedtimes and improved sleep hygiene may more readily prevent sleep deprivation in this student group. Nevertheless, if parents do not alter their children's sleep behavior in response to earlier start times, elementary school performance may suffer, and these reductions in early student learning may have implications for academic achievement over the long term (G. W. Ladd & Dinella, 2009). On the other hand, making school start times later for all grade levels may be a feasible solution for some school districts (Kirby et al., 2011).

The association between later start times and higher retention rates was unexpected and indicates that later school start times were associated with a greater number of children being held back a grade. To our knowledge, this is the first study to examine student retention in relation to school start times, and it is therefore difficult to draw firm conclusions about this finding. However, given that other indices of school performance were improved at later school start times, one possible explanation is that once the average students begin to improve, students with learning difficulties have an especially hard time keeping up. Lagging further behind the majority of students may lead to retention. This explanation is somewhat consistent with the findings that later school start times tend to benefit only those schools that have more middle or upper class students. On the other hand, this finding is inconsistent with other research suggesting that students with the lowest scores benefit from later school start times the most (Edwards, 2012).

Appalachian county designation did not moderate any associations, although it was consistently related to poorer school performance. On the other hand, the percentage of students qualifying for free and reduced-cost lunch (based on family income and therefore a measure of low socioeconomic status) consistently moderated associations between school start times and school academic success. Significant relations between early school start times and poor school performance were found only for schools with a lower percentage of students qualifying for free and reduced-cost lunches (e.g., for schools with a wealthier student population). In other words, schools with economically disadvantaged students were unlikely to show better school performance if their start times were later. This is inconsistent with recent policy proposals suggesting that later school start times are a promising mechanism for closing the achievement gap between poor and wealthy students (Jacob & Rockoff, 2011).

This lack of improvement in poorer school systems may be explained through a cumulative risk model (Evans, 2004; Sameroff, Seifer, Barcas, Zax, & Greenspan, 1987). According to Dubow and Ippolito (1994), poverty may be one of the single

greatest risk factors for student academic performance. According to the cumulative risk model, poverty influences child development because of the accumulation of multiple stressors that accompany poverty (Sameroff et al., 1987). Indeed, poverty has been linked to a wide range of stressors in both the psychosocial and physical environments (Evans, 2004). For example, the psychosocial environment of poverty may be characterized by exposure to violence (Emery & Laumann-Billings, 1998), marital conflict or divorce (Liu & Chen, 2006), harsh and unresponsive parenting (Conger & Elder, 1994; Grant et al., 2003), low parental monitoring (Kilgore, Snyder, & Lentz, 2000), less cognitive stimulation (Hoff, Laursen, & Tardiff, 2002), less parental involvement in school systems (Benveniste, Carnoy, & Rothstein, 2003), schools with less highly trained teachers and greater violence (Clotfelter, Ladd, Vigdor, & Wheeler, 2006; Milam, Furr-Holden, & Leaf, 2010), and changes in schools and residences (Herbers et al., 2012). The physical environment of poverty may be characterized by exposure to toxins and parental smoking (Centers for Disease Control and Prevention, 2010; Legot, London, Rosofsky, & Sandra, 2012), noise (Evans & Kim, 2012), crowded housing conditions (Myers, Baer, & Choi, 1996), inadequate heat (Children's Defense Fund, 1995), lack of air conditioning (Federman et al., 1996), poor nutrition (Alaimo, Olson, Frongillo, & Briefel, 2001), and crumbling schools (National Center for Education Statistics, 2000).

The cumulative model of risk posits that no one specific risk factor is tied to child developmental outcomes. Rather, it is the number of risk factors that predict developmental outcomes, including allostatic load, academic achievement, and mental health (Appleyard, Egeland, van Dulmen, & Sroufe, 2005). Several studies now indicate that the presence of four or more risk factors conveys special risk for compromised development (Sameroff, Bartko, Baldwin, Baldwin, & Seifer, 1998). Children growing up in poverty are likely to experience this number of risks. Low income fourth graders have 35% more negative life events in a year than middle income fourth graders (Attar, Guerra, & Tolan, 1994). Other studies report even larger discrepancies based on income; approximately 35% of children living in poverty—compared to only 5% in wealthier families—have six or more risk factors present in their lives (Liaw & Brooks-Gunn, 1994). The increased risk burden mediates the association between poverty and psychophysiological functioning and psychological stress (Evans & English, 2002).

The implication is that removing one risk factor may have little impact, unless it brings the child under the risk threshold. At the same time, there is an incremental influence over time: The longer one is exposed to the stresses and disadvantages associated with poverty, the greater the risk and the poorer the outcomes in psychological and cognitive domains (Lynch, Kaplan, & Shema, 1997). The impact of later school start times for impoverished school children may therefore be too little, too late, for academic performance. Indeed, later school start times may not even improve sleep in poor children. There is an increased incidence of sleep problems in the context of poverty, perhaps because of less comfortable sleep surfaces and room temperatures, room sharing, noise, and poor sleep hygiene (Buckhalt & Staton, 2011). As such, a delay in school start times may not be sufficient to overcome the numerous other obstacles that children in poverty face, including obstacles to obtaining adequate sleep.

Limitations

The current study did not assess sleep directly and did not differentiate different aspects of sleep. A meta-analysis about sleep and school performance has shown that different measures of sleep condition are related to school performance to differing extents: Sleepiness is most strongly related to school performance, followed by sleep quality and sleep duration (Dewald et al., 2010). Earlier school start time may jeopardize different facets of sleep, and further research is needed to differentiate these. The current study is also limited by its cross-sectional design and data from only one state. Although we controlled for a number of potential confounding factors, including the racial composition of the schools and teacher-student ratio, we cannot infer that early school start times were the cause of school performance measures. Findings may not generalize to other states, especially to states that have varying levels of poverty or more racial diversity than Kentucky. Finally, we used traditional estimation methods to predict school rank; this variable is a rank order variable, and the traditional estimation procedure may yield somewhat inaccurate estimates.

Despite these limitations, this study addresses some key gaps in the current literature on school start times. First, we demonstrate that there are associations between early school start times and school performance, particularly among elementary schools serving middle and upper class students. Identifying school characteristics that moderate associations between school start times and school performance has rarely been done for this topic. Finally, we provide one of the very few examinations of school start times and test scores in elementary schools. Our findings indicate that early school start times may be just as detrimental for young children as they are for adolescents.

References

Alaimo, K., Olson, C. M., Frongillo, E. A., & Briefel, R. R. (2001). Food insufficiency, family income, and health in U.S. preschool and school-aged children. *American Journal of Public Health, 91*, 781–786. doi: 10.2105/AJPH.91.5.781

Anderson, B., Storfer-Isser, A., Taylor, H. G., Rosen, C. L., & Redline, S. (2009). Associations of executive function with sleepiness and sleep duration in adolescents. *Pediatrics, 123*, 701–707. doi:10.1542/peds.2008-1182

Appleyard, K., Egeland, B., van Dulmen, M. H. M., & Sroufe, L. A. (2005). When more is not better: The role of cumulative risk in child behavior outcomes. *Journal of Child Psychology and Psychiatry, 46*, 235–245. doi:10.1111/j.1469-7610.2004.00351.x

Astill, R. G., Van der Heijden, K. B., Van IJzendoorn, M. H., & Van Someren, E. J. (2012). Sleep, cognition, and behavioral problems in school-age children: A century of research meta-analyzed. *Psychological Bulletin, 138*, 1109–1138. doi:10.1037/a0028204

Attar, B., Guerra, N., & Tolan, P. (1994). Neighborhood disadvantage, stressful life events, and adjustment in urban elementary school children. *Journal of Clinical Child Psychology, 23*, 391–400. doi:10.1207/s15374424jccp2304_5

Bauer, D. J., & Sterba, S. K. (2011). Fitting multilevel models with ordinal outcomes: Performance of alternative specifications and methods of estimation. *Psychological Methods, 16*, 373–390. doi:10.1037/a0025813

Beebe, D. W. (2011). Cognitive, behavioral, and functional consequences of inadequate sleep in children and adolescents. *Pediatric Clinics of North America, 58*, 649–665. doi:10.1016/j.pcl.2011.03.002

Benveniste, L., Carnoy, M., & Rothstein, R. (2003). *All else equal*. New York, NY: Routledge-Farmer.

Bryant, P. A., Trinder, J., & Curtis, N. (2004). Sick and tired: Does sleep have a vital role in the immune system? *Nature Reviews Immunology, 4*, 457–467. doi:10.1038/nri1369

Buckhalt, J. A. (2011). Insufficient sleep and the socioeconomic status achievement gap. *Child Development Perspectives, 5*, 59–65. doi: 10.1111/j.1750-8606.2010.00151.x

Buckhalt, J. A., El-Sheikh, M., Keller, P. S., & Kelly, R. J. (2009). Concurrent and longitudinal relations between children's sleep and cognitive functioning: The moderating role of parent education. *Child Development, 80*, 875–892. doi:10.1111/j.1467-8624.2009.01303.x19489909

Buckhalt, J. A., & Staton, L. E. (2011). Children's sleep, cognition, and academic performance in the context of socioeconomic status and ethnicity. In M. El-Sheikh (Ed.), *Sleep and development: Familial and socio-cultural considerations* (pp. 245–264). New York, NY: Oxford University Press. doi:10.1093/acprof:oso/9780195395754.003.0011

Cappuccio, F. P., Cooper, D., D'Elia, L., Strazzullo, P., & Miller, M. A. (2011). Sleep duration predicts cardiovascular outcomes: A systematic review and meta-analysis of prospective studies. *European Heart Journal, 32*, 1484–1492. doi:10.1093/eurheartj/ehr007

Carskadon, M. A., Wolfson, A. R., Acebo, C., Tzischinsky, O., & Seifer, R. (1998). Adolescent sleep patterns, circadian timing, and sleepiness at a transition to early school days. *Sleep: Journal of Sleep Research & Sleep Medicine, 21*, 871–881.

Centers for Disease Control and Prevention. (2010). Vital signs: Current cigarette smoking among adults aged ≥ 18 years—United States, 2009. *Morbidity and Mortality Weekly Report, 59*, 1135–1140.

Chen, C. L., Liu, T. T., Yi, C. H., & Orr, W. C. (2011). Evidence for altered anorectal function in irritable bowel syndrome patients with sleep disturbance. *Digestion, 84*, 247–251. doi:10.1159/000330847

Chervin, R. D., Ruzicka, D. L., Giordani, B. J., Weatherly, R. A., Dillon, J. E., Hedges, E. K., . . . Guire, K. E. (2006). Sleep-disordered breathing, behavior, and cognition in children before and after adenotonsillectomy. *Pediatrics, 117*, 769–778. doi:10.1542/peds.2005-1837

Children's Defense Fund. (1995). *The state of America's children yearbook 1995*. Washington, DC: Author

Clotfelter, C., Ladd, H. F., Vigdor, J., & Wheeler, J. (2006). High-poverty schools and the distribution of teachers and principals. *North Carolina Law Review, 85*, 1345–1379.

Conger, R. D., & Elder, G. H. (1994). *Families in troubled times*. New York, NY: Aldine de Gruyter.

Crowley, S. J., Acebo, C., & Carskadon, M. A. (2007). Sleep, circadian rhythms, and delayed phase in adolescence. *Sleep Medicine, 8*, 602–612. doi:10.1016/j.sleep.2006.12.002

Curcio, G., Ferrara, M., & De Gennaro, L. (2006). Sleep loss, learning capacity and academic performance. *Sleep Medicine Reviews, 10*, 323–337. doi:10.1016/j.smrv.2005.11.001

Dettmers, S., Trautwein, U., Lüdtke, O., Kunter, M., & Baumert, J. (2010). Homework works if homework quality is high: Using multilevel modeling to predict the development of achievement in mathematics. *Journal of Educational Psychology, 102*, 467–482. doi:10.1037/a0018453

Dewald, J. F., Meijer, A. M., Oort, F. J., Kerkhof, G. A., & Bogels, S. M. (2010). The influence of sleep quality, sleep duration and sleepiness on school performance in children and adolescents: A meta-analytic review. *Sleep Medicine Reviews, 14*, 179–189. doi:10.1016/j.smrv.2009.10.004

Dexter, D., Bijwadaria, J., Schilling, D., & Applebaugh, G. (2003). Sleep, sleepiness and school start times: A preliminary study. *Wisconsin Medical Journal, 102*(1), 44–46.

de Young, A. J. (1985). Economic-development and educational status in Appalachian Kentucky. *Comparative Education Review, 29*(1), 47–67. doi:10.1086/446488

Dubow, E. F., & Ippolito, M. F. (1994). Effects of poverty and quality of the home environment on changes in the academic and behavioral adjustment of elementary school-age children. *Journal of Clinical Child Psychology*, 23, 401–412. doi:10.1207/s15374424jccp2304_6

Durmer, J. S., & Dinges, D. F. (2005). Neurocognitive consequences of sleep deprivation. *Seminars in Neurology*, 25, 117–129. doi:10.1055/s-2005-867080

Dworak, M., Schierl, T., Bruns, T., & Struder, H. K. (2007). Impact of singular excessive computer game and television exposure on sleep patterns and memory performance of school-aged children. *Pediatrics*, 120, 978–985. doi:10.1542/peds.2007-0476

Edwards, F. (2012). Early to rise? The effect of daily start times on academic performance. *Economics of Education Review*, 31, 970–983. doi:10.1016/j.econedurev.2012.07.006

El-Sheikh, M., Buckhalt, J. A., Keller, P. S., Cummings, E. M., & Acebo, C. (2007). Child emotional insecurity and academic achievement: The role of sleep disruptions. *Journal of Family Psychology*, 21, 29–38. doi:10.1037/0893-3200.21.1.29

Emery, R. E., & Laumann-Billings, L. (1998). An overview of the nature, causes and consequences of abusive family relationships. *American Psychologist*, 53, 121–135. doi:10.1037/0003-066X.53.2.121

Epstein, R., Chillag, N., & Lavie, P. (1998). Starting times of school: Effects on daytime functioning of fifth-grade children in Israel. *Sleep: Journal of Sleep Research & Sleep Medicine*, 21, 250–256.

Evans, G. W. (2004). The environment of childhood poverty. *American Psychologist*, 59, 77–92. doi:10.1037/0003-066X.59.2.77

Evans, G. W., & English, K. (2002). The environment of poverty: Multiple stressor exposure, psychophysiological stress, and socioemotional adjustment. *Child Development*, 73, 1238–1248. doi:10.1111/1467-8624.00469

Evans, G. W., & Kim, P. (2012). Childhood poverty and young adults' allostatic load: The mediating role of childhood cumulative risk exposure. *Psychological Science*, 23, 979–983. doi:10.1177/0956797612441218

Fallone, G., Owens, J. A., & Deane, J. (2002). Sleepiness in children and adolescents: Clinical implications. *Sleep Medicine Reviews*, 6, 287–306. doi:10.1053/smrv.2001.0192

Federman, M., Garner, T., Short, K., Cutter, W., Levine, D., McGough, D., & McMillen, M. (1996, May). What does it mean to be poor in America? *Monthly Labor Review*, (5), 3–17.

Goddard, R. D., & Goddard, Y. L. (2001). A multilevel analysis of the relationship between teacher and collective efficacy in urban schools. *Teaching and Teacher Education*, 17, 807–818. doi:10.1016/S0742-051X(01)00032-4

Grant, K. E., Compas, B. E., Stuhlmacher, A., Thurm, A., McMahon, S., & Halpert, J. (2003). Stressors and child and adolescent psychopathology: Moving from markers to mechanisms of risk. *Psychological Bulletin*, 129, 447–466. doi:10.1037/0033-295X.129.3.447

Herbers, J. E., Cutuli, J. J., Supkoff, L. M., Heistad, D., Chan, C. K., Hinz, E., & Masten, A. S. (2012). Early reading skills and academic achievement trajectories of students facing poverty, homelessness, and high residential mobility. *Educational Researcher*, 41, 366–374. doi:10.3102/0013189X12445320

Hoff, E., Laursen, B., & Tardiff, T. (2002). Socioeconomic status and parenting. In M. H. Bornstein (Ed.), *Handbook of parenting* (2nd ed., pp. 231–252). Mahwah, NJ: Erlbaum.

Irwin, M., McClintick, J., Costlow, C., Fortner, M., White, J., & Gillin, J. C. (1996). Partial night sleep deprivation reduces natural killer and cellular immune responses in humans. *FASEB Journal*, 10, 643–653.

Jacob, B. A., & Rockoff, J. E. (2011). *Organizing schools to improve student achievement: Start times, grade configurations, and teacher assignments*. Washington, DC: Hamilton Project.

Kakkar, R. K., & Berry, R. B. (2009). Asthma and obstructive sleep apnea: At different ends of the same airway? *Chest*, 135, 1115–1116. doi:10.1378/chest.08-2778

Kentucky Department of Education. (2011). *Kentucky school report cards (2011–2012)* [Data set]. Retrieved from <http://applications.education.ky.gov/SRC/DataSets.aspx>

Kentucky Department of Education. (2012). Researchers. Retrieved from <http://education.ky.gov/Pages/default.aspx>

Kilgore, K., Snyder, J., & Lentz, C. (2000). The contribution of parental discipline, parental monitoring, and school risk to early-onset conduct problems in African American boys and girls. *Developmental Psychology*, 36, 835–845. doi:10.1037/0012-1649.36.6.835

Kirby, M., Maggi, S., & D'Angiulli, A. (2011). School start times and the sleep–wake cycle of adolescents: A review and critical evaluation of available evidence. *Educational Researcher*, 40, 56–61. doi:10.3102/0013189X11402323

Knutson, K. L., & Lauderdale, D. S. (2009). Sociodemographic and behavioral predictors of bed time and wake time among U.S. adolescents aged 15 to 17 years. *The Journal of Pediatrics*, 154, 426–430. doi:10.1016/j.jpeds.2008.08.035

Kopasz, M., Loessl, B., Hornyak, M., Riemann, D., Nissen, C., Piosczyk, H., & Voderholzer, U. (2010). Sleep and memory in healthy children and adolescents – a critical review. *Sleep Medicine Reviews*, 14, 167–177. doi:10.1016/j.smrv.2009.10.006

Ladd, G. W., & Dinella, L. M. (2009). Continuity and change in early school engagement: Predictive of children's achievement trajectories from first to eighth grade? *Journal of Educational Psychology*, 101, 190–206. doi:10.1037/a0013153

Ladd, H. F. (2012). Education and poverty: Confronting the evidence. *Journal of Policy Analysis and Management*, 31, 203–227. doi:10.1002/pam.21615

Laird, J., Cataldi, E. F., KewalRamani, A., & Chaoman, C. (2008). *Dropout and completion rates in the United States: 2006* (Report No. 2008-053). Retrieved from <http://nces.edu/pubsearch/pubsinfo.asp?pubid=2008053>

Legot, C., London, B., Rosofsky, A., & Shandra, J. (2012). Proximity to industrial toxins and childhood respiratory, developmental, and neurological diseases: Environmental ascription in East Baton Rouge Parish, Louisiana. *Population and Environment*, 33, 333–346. doi:10.1007/s11111-011-0147-z

Li, S. H., Arguelles, L., Jiang, F., Chen, W. J., Jin, X. M., Yan, C. H., . . . Shen, X. M. (2013). Sleep, school performance, and a school-based intervention among school-aged children: A sleep series study in China. *PLOS One*, 8(7), e67928. doi:10.1371/journal.pone.0067928

Liaw, F., & Brooks-Gunn, J. (1994). Cumulative familial risks and low birth weight children's cognitive and behavioral development. *Journal of Clinical Child Psychology*, 23, 360–372. doi:10.1207/s15374424jccp2304_2

Liu, R. X., & Chen, Z. (2006). The effects of marital conflict and marital disruption on depressive affect: A comparison between women in and out of poverty. *Social Science Quarterly*, 87, 250–271. doi:10.1111/j.1540-6237.2006.00379.x

Lufi, D., Tzischinsky, O., & Hadar, S. (2011). Delaying school starting time by one hour: Some effects on attention levels in adolescents. *Journal of Clinical Sleep Medicine*, 7, 137–143.

Lynch, J. W., Kaplan, G. A., & Shema, S. J. (1997). Cumulative impact of sustained economic hardship on physical, cognitive, psychological, and social functioning. *The New England Journal of Medicine*, 337, 1889–1895. doi:10.1056/NEJM199712253372606

Milam, A. J., Furr-Holden, C. D. M., & Leaf, P. J. (2010). Perceived school and neighborhood safety, neighborhood violence, and academic achievement in urban school children. *The Urban Review*, 42, 458–467. doi:10.1007/s11256-010-0165-7

Myers, D., Baer, W., & Choi, S. (1996). The changing problems of overcrowded housing. *Journal of the American Planning Association*, 62, 66–84. doi:10.1080/01944369608975671

National Center for Education Statistics. (2000). *Condition of America's public school facilities: 1999* (Report No. 2000–032). Washington, DC: U.S. Department of Education.

National Sleep Foundation. (2005a). *Changing school start times: Fayette County, Kentucky*. Retrieved from <http://www.sleepinfairfax.org/docs/CS.Fayette.pdf>

National Sleep Foundation. (2005b). *Changing school start times: Jessamine County, Kentucky*. Retrieved from <http://www.sleepinfairfax.org/docs/CS.Jessamine.pdf>

National Sleep Foundation. (2011). *2011 Sleep in America poll: Communications technology in the bedroom*. Retrieved from <http://teensneedssleep.files.wordpress.com/2011/05/national-sleep-foundation-2011-sleep-in-america-poll-communications-technology-in-the-bedroom.pdf>

Owens, J. A., Belon, K., & Moss, P. (2010). Impact of delaying school start time on adolescent sleep, mood, and behavior. *Archives of Pediatric and Adolescent Medicine*, 164, 608–614. doi:10.1001/archpediatrics.2010.96

Owens, J., Maxim, R., McGuinn, M., Nobile, C., Msall, M., & Alario, A. (1999). Television-viewing habits and sleep disturbance in school children. *Pediatrics*, 104(3), e27. doi:10.1542/peds.104.3.e27

Paavonen, E. J., Aronen, E. T., Moilanen, I., Piha, J., Rasanen, E., Tamminen, T., & Almqvist, F. (2000). Sleep problems of school-aged children: A complementary view. *Acta Paediatrica*, 89, 223–228. doi:10.1111/j.1651-2227.2000.tb01220.x

Preacher, K. J., Curran, P. J., & Bauer, D. J. (2006). Computational tools for probing interaction effects in multiple linear regression, multilevel modeling, and latent curve analysis. *Journal of Educational and Behavioral Statistics*, 31, 437–448. doi:10.3102/10769986031004437

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis* (2nd ed.). Thousand Oaks, CA: Sage.

Sadeh, A., Gruber, R., & Raviv, A. (2003). The effects of sleep restriction and extension on school-age children: What a difference an hour makes. *Child Development*, 74, 444–455. doi:10.1111/1467-8624.7402008

Sameroff, A. J., Bartko, W. T., Baldwin, A., Baldwin, C., & Seifer, R. (1998). Family and social influences on the development of child competence. In M. Lewis & C. Feiring (Eds.), *Families, risk, and competence* (pp. 161–183). Mahwah, NJ: Erlbaum.

Sameroff, A. J., Seifer, R., Barocas, R., Zax, M., & Greenspan, S. (1987). Intelligence quotient scores of 4-year-old children: Social environmental risk factors. *Pediatrics*, 79, 343–350.

Shaw, T. C., De Young, A. J., & Rademacher, E. W. (2004). Educational attainment in Appalachia: Growing with the nation, but challenges remain. *Journal of Appalachian Studies*, 10, 307–329.

Shen, J., Leslie, J. M., Spybrook, J. K., & Ma, X. (2012). Are principal background and school processes related to teacher job satisfaction? A multilevel study using schools and staffing survey 2003–04. *American Educational Research Journal*, 49, 200–230. doi:10.3102/0002831211419949

Wahlstrom, K. L. (2002). Accommodating the sleep patterns of adolescents within current educational structures: An uncharted path. In M. A. Carskadon (Ed.), *Adolescent sleep patterns: Biological, social, and psychological influences* (pp. 172–197). New York, NY: Cambridge University Press.

Wenglinsky, H. (2002). The link between teacher classroom practices and student academic performance. *Education Policy Analysis Archives*, 10(12), 1–30.

Wilson, S., & Gore, J. (2009). Appalachian origin moderates the association between school connectedness and GPA: Two exploratory studies. *Journal of Appalachian Studies*, 15, 70–86.

Wolfson, A. R., & Carskadon, M. A. (1998). Sleep schedules and daytime functioning in adolescents. *Child Development*, 69, 875–887. doi:10.1111/j.1467-8624.1998.tb06149.x

Wolfson, A. R., Spaulding, N. L., Dandrow, C., & Baroni, E. M. (2007). Middle school start times: The importance of a good night's sleep for young adolescents. *Behavioral Sleep Medicine*, 5, 194–209. doi:10.1080/15402000701263809

Ziliak, J. P. (2012). The Appalachian Regional Development Act and economic change. In J. P. Ziliak (Ed.), *Appalachian legacy: Economic opportunity after the war on poverty* (pp. 19–44). Washington, DC: Brookings Institution Press.

Received November 1, 2013

Revision received April 4, 2014

Accepted April 9, 2014 ■

Developmental Dynamics Between Children's Externalizing Problems, Task-Avoidant Behavior, and Academic Performance in Early School Years: A 4-Year Follow-Up

Riitta-Leena Metsäpelto, Eija Pakarinen, Noona Kiuru, Anna-Maija Poikkeus,
Marja-Kristiina Lerkkanen, and Jari-Erik Nurmi
University of Jyväskylä

This longitudinal study investigated the associations among children's externalizing problems, task-avoidant behavior, and academic performance in early school years. The participants were 586 children (43% girls, 57% boys). Data pertaining to externalizing problems (teacher ratings) and task-avoidant behaviors (mother and teacher ratings) were gathered, and the children were tested yearly on their academic performance in Grades 1–4. The results were similar for both genders. The analyses supported a mediation model: high externalizing problems in Grades 1 and 2 were linked with low academic performance in Grades 3 and 4 through increases in task-avoidant behavior in Grades 2 and 3. The results also provided evidence for a reversed mediator model: low academic performance in Grades 1 and 2 was associated with high externalizing problems in Grades 3 and 4 via high task avoidance in Grades 2 and 3. These findings emphasize the need to examine externalizing problems, task-avoidant behavior, and academic performance conjointly to understand their developmental dynamics in early school years.

Keywords: externalizing problems, task-avoidant behavior, academic performance, longitudinal study, cross-lagged associations

Externalizing problems and maladaptive achievement behaviors constitute major problems in primary school and compromise students' learning outcomes and adjustment at school. Previous research has shown that externalizing problems are linked to low reading and math attainments (Adams, Snowling, Hennessy, & Kind, 1999), lower cognitive abilities and academic achievement (Bub, McCartney, & Willett, 2007), a higher incidence of repeating a class, and a diminished probability of graduating from high school and attending college (McLeod & Kaiser, 2004). Likewise, maladaptive achievement behavior, as indicated by avoidance of learning tasks and adoption of strategies that interfere with learning (e.g., procrastination), has been found to predict subsequent poor academic performance (Aunola, Nurmi, Niemi, Lerkkanen, & Rasku-Puttonen, 2002; Mägi, Häidkind, & Kikas, 2010; Midgley & Urdan, 1995).

In this study, we drew together and extended previous work on these two lines of research by investigating the cross-lagged associations between externalizing problems, task-avoidant behavior,

and academic performance, using multiwave panel data. Prior research has provided only scant evidence of the linkages between externalizing problems and task-avoidant behavior, and little is known about how they might combine to contribute to children's academic performance. Our first question concerned the extent to which externalizing problems influence children's achievement via task-avoidant behavior. The goal was to increase understanding of the mechanisms through which problem behaviors and achievement strategies in the learning contexts are intertwined to affect the academic outcomes of children in the beginning of their school career. Our second question concerned reversed mediator effects, that is, the extent to which low academic performance increases externalizing problems via task-avoidant behavior. Evidence has linked academic difficulties to increases in externalizing problems, but studies investigating the mediating mechanisms are rare.

Externalizing Problems and Academic Performance

Externalizing problems refer to a broad category of disruptive behaviors, such as aggressiveness, oppositional behavior, conduct problems, hyperactivity, and attention deficit problems (McMahon, 1994). Early signs of externalizing problems, manifested in substantial noncompliance, aggression toward peers, high activity level, and poor regulation of impulses, can be identified already in the toddler and preschool years (Campbell, Shaw, & Gilliom, 2000). Although the average levels of externalizing problems tend to decrease from preschool age to young adulthood (Bongers, Koot, van der Ende, & Verhulst, 2003), individual differences in externalizing problems persist at moderate levels from early to middle childhood (Deater-Deckard, Dodge, Bates, & Pettit, 1998) and through adolescence (Broidy et al., 2003). Re-

This article was published Online First July 14, 2014.

Riitta-Leena Metsäpelto and Eija Pakarinen, Department of Teacher Education, University of Jyväskylä; Noona Kiuru, Department of Psychology, University of Jyväskylä; Anna-Maija Poikkeus and Marja-Kristiina Lerkkanen, Department of Teacher Education, University of Jyväskylä; Jari-Erik Nurmi, Department of Psychology, University of Jyväskylä.

This study was funded by grants from the Academy of Finland (Grants No. 252 304 for 2011–2014 and No. 263 891 for 2013–2015).

Correspondence concerning this article should be addressed to Riitta-Leena Metsäpelto, Department of Teacher Education, P.O. Box 35, 40014 University of Jyväskylä, Jyväskylä, Finland. E-mail: riitta-leena.metsapelto@jyu.fi

search on factors predisposing children to externalizing problems implicates multiple risk factors (Deater-Deckard et al., 1998). They include individual factors such as deficits in self-regulation (Olson et al., 2011) and difficult temperament (Miller-Lewis et al., 2006) or contextual factors such as poverty (Grant et al., 2003), parental use of harsh and punitive discipline (Olson et al., 2011), and rejection by peers (Laird, Jordan, Dodge, Pettit, & Bates, 2001). Externalizing behaviors are also more prevalent among boys than girls in early and middle childhood (Bongers et al., 2003; Leadbeater, Kuperminc, Blatt, & Herzog, 1999).

Externalizing problems in early childhood and school age have been shown to predict various adverse mental health and psychosocial outcomes (Caspi, 2000; Fergusson, Horwood, & Ridder, 2007). In addition, children with externalizing problems often fail to take advantage of learning opportunities in the classroom. For instance, children with symptoms or diagnosis of attention-deficit/hyperactivity disorder are much more likely to exhibit academic impairments in reading, writing, and mathematics than children without such symptoms (Barry, Lyman, & Klinger, 2002; McConaughy, Volpe, Antshel, Gordon, & Eiraldi, 2011; Spira & Fischel, 2005). Conduct and oppositional deficit disorders also co-occur with learning difficulties and academic underachievement, although these associations are less consistent and at least partly accounted for by linkages with attention-deficit/hyperactivity disorder (Frick et al., 1991; Hinshaw & Lee, 2003). Literature on the broadband construct of externalizing problems shows that children and adolescents with externalizing problems exhibit deficits both in general academic competence (Burt & Roisman, 2010; Masten et al., 2005) and in the development of more specific skills, such as reading, writing, and math (Adams et al., 1999; Gresham, Lane, MacMillan, & Bocian, 1999; Hinshaw, 1992; Nelson, Benner, Lane, & Smith, 2004). Academic difficulties may result from disruptive behavior, leading students to overlook vital information and fail to follow teachers' instructions (Atkins, McKay, Talbott, & Arvanitis, 1996). The negative dynamic between behavior and achievement may also manifest as avoidance of tasks or assignments in the classroom. Furthermore, children with externalizing problems have more conflicts with teachers and more negative attitudes in teacher-student relationships than children without behavioral difficulties (Henricsson & Rydell, 2004).

The association between externalizing problems and academic outcomes may also run in the opposite direction, that is, academic achievement affecting externalizing behavior. Masten et al. (2005) pointed out that the emergence of emotional and behavioral problems is related to the failure to accomplish age-salient developmental tasks, such as integration into school and successful acquisition of knowledge and skills. Through normative comparisons with their classmates, children become more aware of their academic progress and standing with respect to abilities (Sutherland, Lewis-Palmer, Stichter, & Morgan, 2008). Students for whom schoolwork is difficult develop negative self-perceptions of ability (Chapman, 1988), and they may feel embarrassment, frustration, and general antagonism toward school, which, in turn, may set in motion defiance and aggressive behavior (Miles & Stipek, 2006; Roeser, Eccles, & Strobel, 1998). Accordingly, Halonen, Aunola, Ahonen, and Nurmi (2006) found that problems in learning to read predicted an increase in externalizing problem behavior during the first 2 years of primary school. Moreover, McGee, Williams, Share, Anderson, and Silva (1986) followed a group of boys from

ages 5 to 11 years and found evidence of reciprocal relations with behavior problems predicting reading disability and reading difficulties further aggravating the existing problem behaviors. These kinds of reciprocal effects have also been found among older students. In their follow-up of children from fifth to ninth grade, Zimmermann, Schütte, Taskinen, and Köller (2013) found evidence of an escalating cycle of negative outcomes in that high externalizing problems predicted lower school grades in reading and language skills and math that, in turn, contributed to increased externalizing problems both directly and via lowered self-esteem.

Task-Avoidant Behavior and Academic Performance

In addition to externalizing problems, another frequent concern in schools is students' failure to develop academic skills and knowledge due to a tendency to avoid challenging tasks instead of actively attempting to perform them. This maladaptive achievement strategy has been described using various concepts such as self-handicapping (Jones & Berglas, 1978), helplessness (Dweck & Leggett, 1988), and task-avoidant behavior (Nurmi, Aunola, Salmela-Aro, & Lindroos, 2003; Zhang, Nurmi, Kiuru, Lerkkanen, & Aunola, 2011). These concepts share the key idea that failures in learning situations create a negative self-concept and low efficacy beliefs, which increase the likelihood of developing expectations of future failure, leading to low effort and task avoidance in learning settings (Aunola et al., 2002; Sideridis, 2003). The avoidance of learning tasks has been found to be more common among boys than girls (Midgley & Urdan, 1995; Onatsu-Arvilommi & Nurmi, 2000; Pakarinen et al., 2011).

Task avoidance has been linked to a host of negative consequences, such as slow development in reading and math in early school years (Aunola et al., 2002; Georgiou, Manolitsis, Nurmi, & Parrila, 2010; Hirvonen, Tolvanen, Aunola, & Nurmi, 2012; Mägi et al., 2010), learning difficulties (Sideridis, 2003), and low academic performance in young adulthood (Zuckerman, Kieffer, & Knee, 1998). Conversely, children's ability to focus on tasks, sustain effort, and persist in the face of difficulties has been found to predict better achievement outcomes (Duncan et al., 2007; Hughes, Luo, Kwok, & Loyd, 2008).

The evidence also suggests the opposite predictive path where learning difficulties and slow academic progress predict decreasing task involvement and high avoidance behavior in early school years (Aunola et al., 2002) or already around the transition to primary school (Lepola, Salonen, & Vauras, 2000; Pakarinen et al., 2011). Onatsu-Arvilommi and Nurmi (2000), for instance, reported cumulative cycles in which learning difficulties and task-avoidant behaviors reciprocally influence each other by showing that first graders' tendency to avoid learning tasks decreased their subsequent progress in reading skills. Further, a low level of literacy skills increased their subsequent task-avoidant behaviors.

Developmental Links Among Externalizing Problems, Task-Avoidance, and Academic Performance

Externalizing problems and task-avoidant behavior have typically been investigated separately and little is known about how these risk behaviors co-vary and operate in conjunction with each other to predict academic achievement. In the current research, we aimed to draw these two approaches together by investigating the

cross-lagged associations between externalizing problems and task-avoidant behavior, and their associations with academic performance in the early school years. Developmental dynamics of this kind can only be detected using longitudinal data that allow the controlling of pre-existing and concurrent associations, so that the influences from one domain of functioning to another can be evaluated from a developmental perspective (e.g., Bornstein, Hahn, & Haynes, 2010; Masten et al., 2005).

The sparse available research supports the view that externalizing problems co-occur with low involvement in learning tasks (Arnold, 1997; Coie & Dodge, 1988) and school engagement (Risi, Gerhardstein, & Kistner, 2003; Wagner & Cameto, 2004), significantly increasing the risk for learning difficulties and school dropout. The specific interest in the present study was in examining a mediation model wherein externalizing problems are linked with low academic performance through increases in task-avoidant behavior. There are at least two mechanisms through which externalizing problems may bring out task-avoidant behavior. First, externalizing problems are known to increase negative feedback and conflictual interactions with teachers (Henriksson & Rydell, 2004; Ladd, Birch, & Buhs, 1999; Murray & Murray, 2004; Nurmi, 2012; Stipek & Miles, 2008). The accumulation of unrewarding experiences in classroom interactions and learning situations may generate low competence beliefs, failure expectations, and feelings of animosity toward school that, in turn, lead to low inclination to exert effort in academic work. The links between externalizing symptoms and low task focus have been indicated, for instance, by findings of Coie and Dodge (1988), documenting that aggressive first- and third grade children were more likely to spend time "off task" in the classroom than children with average peer status, and by those of Arnold (1997) showing that the aggressive, hostile, and noncompliant behavior of 4- to 6-year-old boys was associated with low on-task behavior.

Second, externalizing problems are often accompanied by difficulties in attending to and complying with teachers' instructions in learning tasks. Prior research has underlined the central role of early self-regulatory mechanisms, such as effortful control and behavioral inhibition, in the development and persistence of externalizing behavior disorders (Olson, Sameroff, Kerr, Lopez, & Wellman, 2005; Olson et al., 2011). Low skills in controlling attention and behavior (high distractibility) have been shown to predict young children's active task avoidance at school (Hirvonen, Aunola, Alatupa, Viljaranta, & Nurmi, 2013). Thus, deficits in attention and self-regulation of behavior—the key features of externalizing problems—can be expected to lead to low persistence and completion of tasks in learning settings.

As outlined previously, both task-avoidant behavior and externalizing problems disrupt classroom learning processes because they interfere with the child's ability to direct and sustain attention on academic activities and work in a self-regulated fashion. Task-avoidant behavior represents a maladaptive achievement strategy that students use to cope with situational demands and stress when confronted with challenging learning tasks (Nurmi et al., 2003; Zhang et al., 2011). Task avoidance is assumed to be gradually built over time based on a history of academic difficulties and ensuing negative self-perceptions, which lead to expectations of subsequent failure and anxiety in new learning situations. To cope with such expectations and feelings, students use task-avoidant behavior either to decrease anxiousness (Miller, 1987) or to create

an excuse (Jones & Berglas, 1978). A persistent pattern of task avoidance is likely to decrease the time that a child spends in effective academic endeavors and affects his or her choices of learning tasks. In contrast, externalizing problems are not restricted specifically to learning tasks but contain a larger set of negative reactions or out-of-bounds behavior expressed in several kinds of environments (i.e., in academic tasks as well as interpersonal relationships with teachers and peers; McMahon, 1994). When examined simultaneously, task avoidance and externalizing problems appear to have differential associations with academic outcome measures. Using six longitudinal data sets at school entry, Duncan et al. (2007; see also Morgan, Farkas, Tufis, & Sperling, 2008) showed that a child's ability to control and sustain attention and participate in classroom activities—not externalizing problems—predicted later reading and math skills. These findings challenge the body of evidence showing linkages between externalizing problems and academic underachievement (Adams et al., 1999; Hinshaw, 1992; Masten et al., 2005). Therefore, research allowing analysis of mediator effects are needed for researchers to gain understanding of the interplay between externalizing problems and task-avoidance in predicting academic achievement.

The present study tests the assumption that externalizing problems (i.e., poor conduct, hyperactivity, and inattentiveness) affect the child's achievement strategies by increasing task-avoidant behavior, further, leading to poor academic performance. While some of the links that we examine have been documented in prior studies, the designs have typically allowed for investigating only two of the three critical measures without integrating all of them into the same study. A design looking at the mediating paths conjointly allowed us to obtain a more comprehensive picture of how behavioral problems interfere with young students' daily functioning in school and compromise their ability to benefit optimally from the learning opportunities. The examination of mediated effects is valuable because more efficient interventions focusing on the relevant variables in the mediating process can be developed and put into action (MacKinnon & Fairchild, 2009). Prior research on the links between externalizing problems and task-avoidant behavior has also relied on cross-sectional data (Coie & Dodge, 1988) or samples of very young children from low-income families (Arnold, 1997). The present study addresses the need to examine how different developmental domains (e.g., behavioral problems, achievement strategies, and academic achievement) influence each other over longer periods as children move through the school system (Roeser et al., 1998).

In regard to the possibility of mediated paths from academic performance via task avoidance to externalizing problems over time, prior research indicates that early learning problems often set in motion negative developmental cycles (e.g., Aunola, Leskinen, Lerkkanen, & Nurmi, 2004). Consistent associations have been documented between learning difficulties and maladaptive achievement strategies, such as task-avoidant behavior (Halonen et al., 2006; Lepola et al., 2000; McGee et al., 1986; Onatsu-Arvilommi & Nurmi, 2000; Pakarinen et al., 2011). On the basis of these findings, it is plausible that early academic difficulties are related to subsequent task-avoidant behavior, which in turn increases externalizing problems. This association has rarely been tested with the exception of the study by Arnold (1997), which provided evidence of the negative escalating cycle in preschool-age high-risk boys, showing that low academic skills explained

low on-task behavior, which in turn increased externalizing problems.

The Present Study

The purpose of this study was to examine cross-lagged associations between externalizing problems, task-avoidant behavior, and academic performance in the early primary school years. We had the following aims and hypotheses:

1. To what extent do externalizing problems predict academic performance, and is this association mediated by task-avoidant behavior? We assumed that high externalizing problems would predict low academic performance (Hypothesis 1a) and that this association would partly be mediated via high task-avoidant behavior (Hypothesis 1b).

2. To what extent does academic performance predict externalizing problems, and does task-avoidant behavior mediate this association? We hypothesized that low academic performance would predict high externalizing problems (Hypothesis 2a) and that high task-avoidant behavior would partly contribute to this association (Hypothesis 2b).

Method

Participants and Procedure

Participants in the present study were part of an extensive follow-up (Lerkkanen et al., 2006) in which a community sample of Finnish-speaking children ($N = 1,880$) were followed up from kindergarten to the end of Grade 4. The follow-up took place in four towns—two in Central Finland, one in Western Finland, and one in Eastern Finland. The children's average age was 85.82 months ($SD = 3.45$ months) at school entry. At the beginning of the study, the children's parents and teachers were asked for their written consent.

The study sample consisted of a more intensively followed subsample of 586 children (43% girls and 57% boys) drawn from the original sample of 1,880 children. This subsample consisted of children identified with risk for reading difficulties at the end of kindergarten ($n = 282$) and randomly selected control children from the same classrooms ($n = 304$). Risk for reading difficulty (RD; for details, see Lerkkanen, Ahonen, & Poikkeus, 2011) was determined on the basis of kindergarten assessment for pre-reading skills (i.e., letter knowledge, phonemic awareness, and rapid automatized naming) and information on the parents' reading difficulties, indicated by either the mother or father self-reporting "mild" or "severe" problems with reading at school age. A child was identified as having a risk for RD if he or she scored at or below the 15th percentile of the total sample in at least two of the measured skill areas, or if the child scored at or below the 15th percentile in one of the skill areas and the parental questionnaire indicated a family risk. It should be noted that the screening took place at an age at which the children had not received formal reading instruction, and it did not include early decoding skills, phoneme blending, or manipulation skills. Since the criterion for risk was set to be lenient, we would expect that only a subgroup of children identified with this early risk would develop difficulties in reading acquisition. From the other participants of the follow-up ($n = 1,690$), a random sample of nonrisk children who did not

meet the risk criteria were also included in the individual follow-up assessment from the first grade onward ($n = 258$). The random selection of the nonrisk sample was carried out from classrooms in a stratified fashion. Due to variation in class size, the number of nonrisk children from different classrooms ranged between one and six, with a median of three children. Target sampling of children for the individual follow-up was necessary to ensure that the data collection demands placed on the teachers were not too heavy.

The sample was representative of the Finnish population (Statistics Finland, 2007). In 7% (general population 6%) of the families, the parents had not been educated beyond comprehensive school (compulsory education up to the completion of Grade 9); 31% (general population 30%) had completed upper secondary education (senior high school or vocational school, Grades 10–12); 36% (general population 35%) had a bachelor's degree or vocational college degree (3-year education at a college or university); and 26% (general population 29%) had a master's degree (5-year university education) or higher.

The data on the students' externalizing problems (teacher ratings) and task-avoidant behaviors (mother and teacher ratings) were gathered in Grade 1 (April 2008; T1), Grade 2 (April 2009; T2), Grade 3 (April 2010; T3), and Grade 4 (April 2011; T4). The children were tested on their academic performance in Grade 1 (April 2008; T1), Grade 2 (April 2009; T2), Grade 3 (April 2010; T3), and Grade 4 (April 2011; T4).

Measures

Task-avoidant behavior. Task-avoidant behavior was assessed in Grades 1–4 by asking the mothers and teachers to evaluate the extent of the child's task-avoidant behavior using the Behavior Strategy Rating Scale (BSR; Onatsu-Arvilommi & Nurmi, 2000) rated on a 5-point scale (1 = *not at all*; 5 = *to a great extent*). The mothers were asked to evaluate their child's behavior in typical homework situations, and teachers were asked to assess their students' typical behavior in learning situations at school. The combination of mother and teacher ratings provided an assessment of task-avoidant behavior across a variety of learning situations, both at school and at home. The following five items were used:

- (a) Does the child have a tendency to find something else to do instead of focusing on the task at hand?
- (b) If the activity or task is not going well, does the child lose his or her focus?
- (c) Does the child give up on tasks easily?
- (d) Does the child actively attempt to solve even difficult situations and tasks? (reversed)
- (e) Does the child demonstrate initiative and persistence in his or her activities and tasks? (reversed).

The correlations between mother and teacher reports on task avoidance have been found to range from .36 (Grade 1) to .48 (Grade 2), indicating moderate convergent validity (see Zhang et al., 2011). A composite score for task-avoidant behavior for each grade was calculated as a mean of mother's and teacher's items. The Cronbach's alphas for the mean scores were .89, .91, .86, and .89 in Grades 1–4, respectively.

Externalizing problems. Externalizing problems were assessed by teacher-ratings in Grades 1–4 using a Finnish version of

the Strengths and Difficulties Questionnaire (SDQ; Goodman, 1997), which has been shown to be a highly valid screening instrument (Goodman, Ford, Simmons, Gatward, & Meltzer, 2000). The SDQ consists of 25 items rated on a 3-point scale (1 = *not true*, 2 = *somewhat true*, and 3 = *certainly true*), producing scales for hyperactivity/inattention, conduct problems, emotional symptoms, peer problems, and prosociality. We used the scales for hyperactivity/inattention (five items; e.g., restless, overactive, cannot stay still for long, thinks things out before acting [reversed]) and conduct problems (five items; e.g., often fights with other children or bullies them, generally obedient, usually does what adults request [reversed]) to measure externalizing problems. One item, "sees tasks through to the end, good attention span (reversed)," was excluded from the Externalizing Problems Scale because it correlated highest with task-avoidance items, and in the exploratory factor analyses, it had moderate loadings on both the hyperactivity/inattention and task-avoidance factors. The composite score for externalizing problems in Grades 1–4 (Time 1–Time 4; T1–T4) was formed as a mean score of the hyperactivity/inattention and conduct problems scales. The Cronbach's alpha reliabilities for the Externalizing Problems composite were .82, .84, .84, and .78, in Grades 1–4, respectively.

Academic performance. Academic performance in Grades 1–4 was assessed using an aggregate constructed on the basis of the children's scores on reading and arithmetic tasks.

Reading skills (decoding and comprehension) were measured using a nationally standardized test battery developed for students between Grade 1 and Grade 6 (ALLU [Reading test for primary school]; Lindeman, 1998). The *decoding* test was a speeded test in which a maximum of 80 items could be attempted within a 2-min time limit. For each item, there was a picture with four words next to it. The child was asked to read the four phonologically similar words and to draw a line to semantically match the picture to the word. The score was derived by calculating the number of correct answers (maximum score 80). Because of the time limit, the score reflects both the child's fluency in reading the stimulus words and his or her accuracy in making the correct choice from among the alternatives. Differentiation between children's rate of reading acquisition in the highly transparent Finnish language requires a speeded test already at the end of Grade 1 because approximately a third of children learn to decode before entering school and tests of reading accuracy without a time limit reach a ceiling very fast (see Lerkkanen, Rasku-Puttonen, Aunola, & Nurmi, 2004). The parallel versions of the test, A and B, were used on alternate years. Alternate-form reliability between Forms A and B was .84. No ceiling effect was evident in Grade 4. The Kuder–Richardson reliability coefficient for the decoding fluency task in Grades 1–4 was .97, .97, .97, and .87, respectively.

The *reading comprehension* test assessed the child's skills in gleaning factual knowledge, concepts, and inferences from text. The children were asked to answer 12 multiple-choice questions based on silently read text. The children received 1 point for each correct answer (maximum score 12). They completed the task at their own pace, but the maximum time allotted was 45 min. This normed test included different texts and multiple choice questions for Grade 1 through Grade 4 so that task difficulty was optimal for each age. The topics of texts were the following: "Judo" (Grade 1), "Guidelines for Gymnastics" (Grade 2), "Operating a Camera" (Grade 3), and "The Light Requirements of Plants" (Grade 4). No

ceiling effect was evident in Grade 4. The Kuder–Richardson reliability coefficient for the reading comprehension task in Grades 1–4 was .85, .80, .75, and .76, respectively.

Arithmetic skills were assessed using a group-administered Basic Arithmetic Test (BAT; Aunola & Räsänen, 2007), which was designed for Grades 1–6. It contains visually presented addition, subtraction, and multiplication problems (total of 28 items) which become more difficult across primary school years. The test is speeded (3-min time limit), and because of this, it remains challenging even for the oldest children. At the first grade, the test consisted of 14 items for addition (e.g., $2 + 1 = ?$ and $3 + 4 + 6 = ?$) and 14 items for subtraction (e.g., $4 - 1 = ?$ and $20 - 2 - 4 = ?$) problems. The items remain the same until fourth grade, after which some of the easiest tasks were replaced with more challenging ones toward the end of the test (e.g., multiplication problems). The test indexes a combination of speed and accuracy of math performance, and its psychometrics have been shown in a number of earlier publications (e.g., Niemi et al., 2011; Zhang, Koponen, Räsänen, Aunola, Lerkkanen, & Nurmi, 2014). The test score was derived by calculating the total number of correct answers (maximum score 28). The Kuder–Richardson reliability coefficient for the task in Grades 1–4 was .84, .86, .87, and .85, respectively.

To calculate the composite score for Academic Performance in Grades 1–4 (T1–T4), the children's test scores on the reading (decoding fluency and reading comprehension) and arithmetic were standardized and their mean score was calculated. The Cronbach's alpha for the Academic Performance composite score was .76, .69, .60, and .66, in Grades 1–4, respectively.

Analysis Strategy

We first estimated a stability model (M_1 without any cross-lagged paths; see Figure 1) in which externalizing problems, task avoidance, and academic performance were predicted by their preceding values across time. The study variables were allowed to correlate with each other at each time point. In the second model (M_2), cross-lagged paths from task-avoidant behaviors to externalizing problems, from academic performance to task-avoidant behaviors, and from academic performance to externalizing problems were estimated. In the third model (M_3), cross-lagged paths from task-avoidant behaviors to academic performance, from externalizing problems to task-avoidant behaviors, and from externalizing problems to academic performance were estimated. In the fourth model (M_4), all cross-lagged paths were estimated. At each step, the Satorra–Bentler scaled chi-square difference test was used to test the improvement of model fit between the competing models.

The model construction was continued by testing whether the cross-lagged paths could be constrained as equal across the grades. The Satorra–Bentler scaled chi-square difference test was used to determine if constraining the cross-lag paths equal at different grades resulted in a better fit than the model in which the paths were allowed to be freely estimated. As the next step, the statistical significance of hypothesized mediated effects was examined. Finally, the multigroup procedure was used to test whether the paths differ between boys and girls or between children at risk for reading difficulties and nonrisk control children.

The analyses were performed with the Mplus statistical package (Version 7; Muthén & Muthén, 1998–2012) using the standard

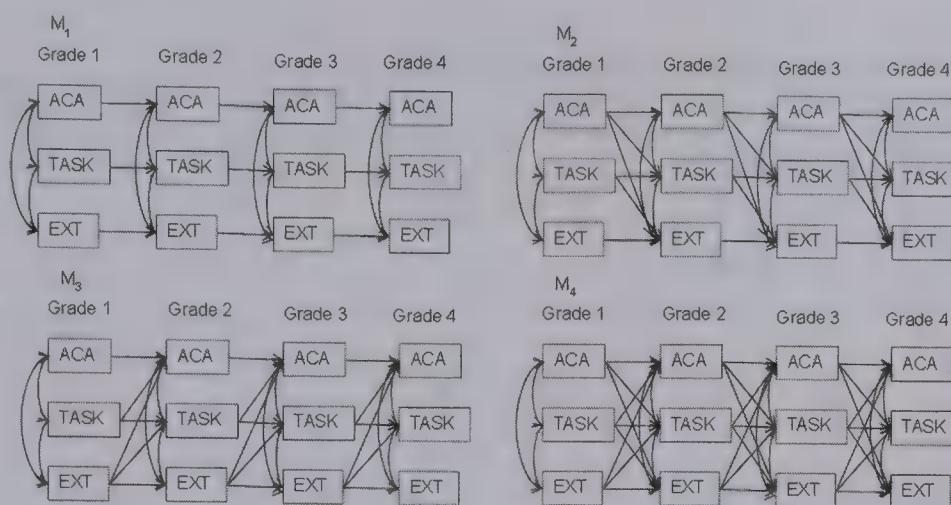


Figure 1. Test of cross-legged relationships in the study with four models (M₁–M₄). ACA = academic performance; TASK = task-avoidant behavior; EXT = externalizing problems.

missing at random (MAR) approach and full-information maximum-likelihood estimation with nonnormality robust standard errors (MLR; Muthén & Muthén, 1998–2012). The goodness of fit of the estimated models was evaluated according to the following four indicators: (a) chi-square test, (b) comparative fit index (CFI), (c) root-mean-square error of approximation (RMSEA), and (d) standardized root-mean-square residual (SRMR).

Since the data were hierarchical in nature (i.e., each teacher assessed more than one child), we investigated the interdependency of the children's ($N = 586$) externalizing problems and task avoidance within classrooms ($N = 155$ classrooms). This investigation was done by means of calculating the intraclass correlations (ICC) using the teacher identification number as a clustering variable (Heck & Thomas, 2009). The resulting ICCs for externalizing problems were .07, .09, .08, and .13, and $ps = .08, .05, .07$, and .03, in Grades 1–4, respectively. The ICCs for teacher-rated task-avoidance were .05, .04, .12, and .12, and $ps = .24, .41, .01$, and .02, in Grades 1–4, respectively. Next, the design effects were obtained according to the ICCs in Grades 1–4, and they were 1.19, 1.21, 1.18, and 1.27 for externalizing problems and 1.13, 1.08, 1.26, and 1.26 for task-avoidance, respectively. Hox and Maas (2002) suggested that analyzing multilevel data as single-level data can yield acceptable (not overly biased) parameter estimates and inferential tests, if the design effects are smaller than 2.0. However, in our subsequent analyses, we used Type = COMPLEX approach (Muthén & Muthén, 1998–2012). This command estimates the model at the whole sample level but corrects for distortions in standard errors in the estimation caused by the clustering of observations (i.e., classroom differences).

Results

Descriptives and Correlations

The descriptive statistics are shown in Table 1. The correlations of study variables across grades indicated moderate-to-high interindividual stability. For externalizing problems, the correlations ranged from .55 to .72, and those for task-avoidant behavior and academic performance ranged from .54 to .72 and from .71 to .81, respectively. Externalizing problems showed moderate positive

correlations with task-avoidant behavior. Furthermore, both externalizing problems and task-avoidant behavior showed moderate or low negative correlations with academic performance.

Cross-Lagged Relationships Among Externalizing Problems, Task-Avoidant Behavior, and Academic Performance

We first estimated the stability model (M₁) without any cross-lagged paths (Figure 1). The goodness-of-fit indices indicated a poor model fit, $\chi^2(46, N = 586) = 451.48, p < .001$; CFI = 0.89, RMSEA = 0.12, SRMR = 0.16. The modification indices suggested that the fit would be improved if there was a direct path (a) from academic performance in Grade 1 to academic performance in Grade 4, (b) from academic performance in Grade 2 to academic performance in Grade 4, (c) from academic performance in Grade 1 to academic performance in Grade 3, (d) from task-avoidant behavior in Grade 1 to task-avoidant behavior in Grade 3, (e) from task-avoidant behavior in Grade 2 to task-avoidant behavior in Grade 4, (f) from externalizing problems in Grade 1 to externalizing problems in Grade 3, and (g) from externalizing problems in Grade 2 to externalizing problems in Grade 4. After adding these paths, the model fit the data well: $\chi^2(40, N = 586) = 188.61, p = .000$; CFI = 0.96, RMSEA = 0.08, SRMR = 0.11.

We then estimated M₂, in which paths from task-avoidant behaviors to externalizing problems, from academic performance to task-avoidant behaviors, and from academic performance to externalizing problems were estimated; M₃ in which paths from task-avoidant behaviors to academic performance, from externalizing problems to task-avoidant behaviors, and from externalizing problems to academic performance were estimated; and finally, M₄, in which all cross-lagged paths were estimated (Figure 1). As can be seen in Table 2, the Satorra–Bentler-scaled chi-square difference test showed that the difference between the stability model (M₁) and M₂ was statistically significant, with M₂ better accounting for the data. The Satorra–Bentler-scaled chi-square difference test further showed that the difference between the stability model M₁ and M₃ was statistically significant, indicating that M₃ provided a better fit with the data. The comparison between stability model M₁ and M₄ showed that the difference was statistically significant,

Table 1
Sample Correlation Matrix and Descriptive Statistics of the Study Variables

Variable	1	2	3	4	5	6	7	8	9	10	11	12
1. Externalizing problems (T1, Grade 1)	1.00											
2. Externalizing problems (T2, Grade 2)	.72	1.00										
3. Externalizing problems (T3, Grade 3)	.60	.68	1.00									
4. Externalizing problems (T4, Grade 4)	.55	.63	.67	1.00								
5. Task avoidance (T1, Grade 1)	.59	.49	.43	.41	1.00							
6. Task avoidance (T2, Grade 2)	.50	.61	.50	.47	.70	1.00						
7. Task avoidance (T3, Grade 3)	.44	.55	.59	.51	.61	.69	1.00					
8. Task avoidance (T4, Grade 4)	.46	.52	.51	.62	.54	.60	.72	1.00				
9. Academic performance (T1, Grade 1)	-.21	-.23	-.18	-.18	-.45	-.43	-.44	-.36	1.00			
10. Academic performance (T2, Grade 2)	-.25	-.28	-.22	-.27	-.44	-.44	-.43	-.39	.81	1.00		
11. Academic performance (T3, Grade 3)	-.24	-.25	-.24	-.25	-.39	-.41	-.44	-.40	.71	.79	1.00	
12. Academic performance (T4, Grade 4)	-.23	-.29	-.27	-.31	-.41	-.45	-.47	-.43	.73	.81	.80	1.00
Mean	1.49	1.47	1.47	1.39	2.67	2.64	2.65	2.53	-0.24	-0.24	-0.21	-0.19
SD	0.47	0.48	0.48	0.40	0.92	0.92	0.94	0.92	0.84	0.84	0.81	0.80
Median	1.33	1.33	1.33	1.22	2.60	2.60	2.60	2.40	-0.30	-0.17	-0.11	-0.14
Min	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-2.05	-2.49	-3.77	-3.16
Max	2.89	2.89	3.00	3.00	5.00	5.00	5.00	5.00	2.76	2.32	1.60	1.83
25 percentile	1.11	1.11	1.11	1.11	2.00	1.90	1.90	1.80	-0.85	-0.82	-0.69	-0.74
75 percentile	1.78	1.78	1.78	1.67	3.37	3.30	3.30	3.20	0.34	0.33	0.39	0.38
Valid <i>N</i>	485	487	474	440	553	546	541	513	586	572	567	543
% missing	17.2	16.9	19.1	24.9	5.8	7.0	7.8	12.6	0	2.4	3.2	7.3

Note. All correlations were significant at $p < .01$ (two-tailed testing of significance). T = Time; Min = minimum; Max = maximum.

indicating that M_4 better accounted for the data. The further model comparisons (M_2 vs. M_4 and M_3 vs. M_4) revealed that M_4 , with all the hypothesized cross-lagged paths, best fit the data.

Next, the cross-lagged paths were set equal across the grades. The Satorra-Bentler scaled χ^2 difference test showed that the model fit was not significantly decreased if all the cross-lagged paths were constrained to be equal across Grades 1–4 ($p > .05$).

The fit of the final model M_4 was $\chi^2(37, N = 586) = 49.46, p = .08$; CFI = 1.00, RMSEA = 0.02, SRMR = 0.02. The results of

this model are presented in Figure 2. High externalizing problems in Grades 1, 2, and 3 predicted high task-avoidant behavior in Grades 2, 3, and 4. High task-avoidant behavior in Grades 2 and 3, in turn, predicted a low academic performance in Grades 3 and 4. However, externalizing problems did not directly predict academic performance at any grade. In addition, the results showed that low academic performance in Grades 1, 2, and 3 predicted high task-avoidance in Grades 2, 3, and 4. High task avoidance in Grades 2 and 3 was associated with higher externalizing problems in Grades 3 and 4. Academic performance was not directly associated with externalizing problems.

The statistical significance of the hypothesized mediator effects was tested. The estimates and standard errors regarding indirect effects are presented in Table 3. The results supported the hypothesized mediator effects. High externalizing problems in Grades 1 and 2 were linked with low academic performance in Grades 3 and 4 through increases in task-avoidant behavior in Grades 2 and 3. Conversely, low academic performance in Grades 1 and 2 was associated with high externalizing problems in Grades 3 and 4 via high task avoidance in Grades 2 and 3.

Additional Analyses

The multigroup method was used to test whether the previous paths differ between boys and girls. The Satorra-Bentler-scaled chi-square difference tests revealed that the model fit was not significantly decreased if the main effects among girls and boys

Table 2
Goodness-of-Fit Statistics (Chi-Square) for the Nested Models

Tested models	χ^2	<i>df</i>	Model comparisons: Satorra-Bentler-scaled χ^2 difference test
1. No cross-lagged paths (M_1)	188.610	40	
2. Cross paths (M_2)	108.086	31	M_1 vs. $M_2 \chi^2(9) = 80.371, p < .001$
3. Cross paths (M_3)	106.086	31	M_1 vs. $M_3 \chi^2(9) = 89.963, p < .001$
4. All cross paths (M_4)	40.226	25	M_1 vs. $M_4 \chi^2(15) = 151.093, p < .001$ M_2 vs. $M_4 \chi^2(6) = 72.611, p < .001$ M_3 vs. $M_4 \chi^2(6) = 61.058, p < .001$

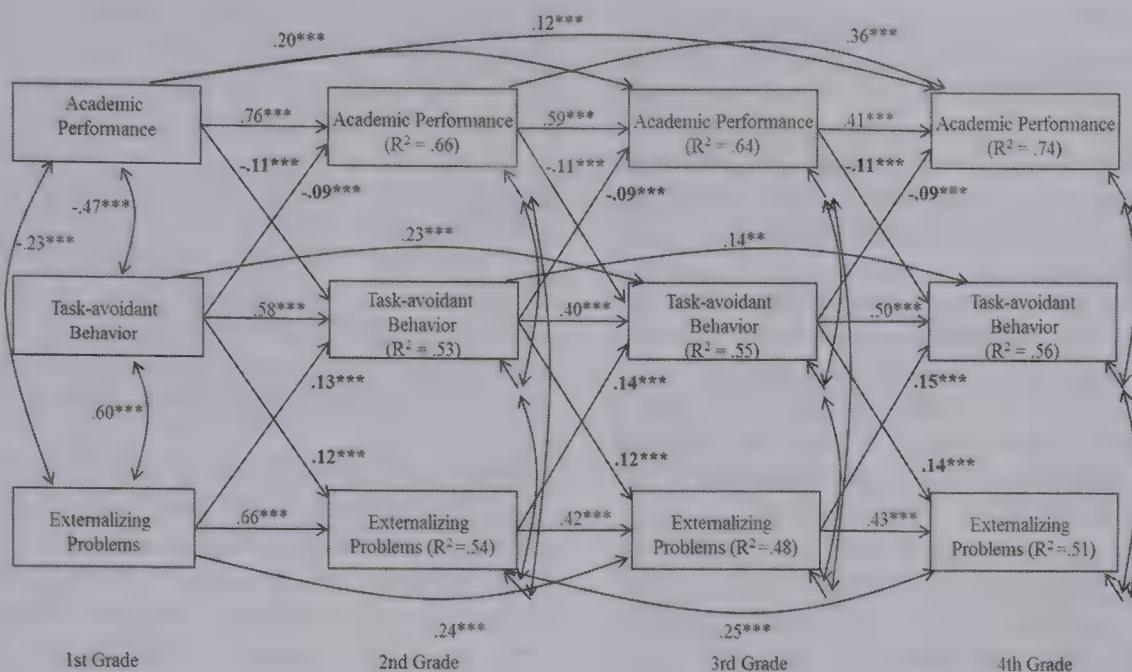


Figure 2. Results of final model M₄: $\chi^2(37) = 49.46$; $p = .08$; comparative fit index = 1.00; root-mean-square error of approximation = 0.02; standardized root-mean-square residual = 0.02. Paths are presented as standardized estimates. Estimates in bold typeface (cross-lagged paths) are constrained to be equal across grades.
 ** $p < .01$. *** $p < .001$.

were constrained to be equal ($p > .05$). Similarly, the multigroup method was used to test whether the previous paths differ between children with a risk of reading difficulties and nonrisk children. The Satorra-Bentler-scaled chi-square difference tests revealed that the model fit was not significantly decreased if the main effects among at-risk children and controls were constrained as equal ($p > .05$). Thus, the multigroup analyses revealed no group differences suggesting that the cross-lagged path models were similar among boys and girls and also among children with and without risk for reading difficulties.

Discussion

In this study, we utilized four-wave cross-lagged panel data to investigate the associations among children's externalizing prob-

lems, task-avoidant behavior, and academic performance in early school years. The results supported a mediation model in which the high externalizing problems in Grades 1 and 2 were linked with low academic performance in Grades 3 and 4 through increases in task-avoidant behavior in Grades 2 and 3. The results also provided evidence for a reversed mediator model: Low academic performance in Grades 1 and 2 was associated with high externalizing problems in Grades 3 and 4 via high task avoidance in Grades 2 and 3. The results were similar for both genders.

The first aim of this study was to investigate the mechanisms through which externalizing problems may impact children's achievement. Instead of direct associations between externalizing problems and academic performance (Hypothesis 1a), we found evidence for an indirect mechanism through which externalizing problems set the stage for increased task-avoidant behavior in learning situations, which leads to lower academic performance (Hypothesis 1b). The indirect linkages were observed beyond autocorrelation, and they were strikingly consistent across different time points of the longitudinal study. These findings are particularly important because our focus was on children who had just entered primary school. This period is critical in the development of basic academic skills and achievement-related strategies that children use to achieve learning goals (Mägi, Lerkkanen, Poikkeus, Rasku-Puttonen, & Kikas, 2010).

There are at least two mechanisms that may be responsible for the mediating role of task-avoidant behavior on the relation between externalizing problems and children's academic performance. First, externalizing problems are likely to increase conflictual interactions and negative feedback received from teachers to the child (Henricksson & Rydell, 2004; Ladd et al., 1999; Stipek & Miles, 2008), which, in turn, may generate low competence beliefs and failure expectations, and finally low inclination to exert effort needed for success in academic work (Nurmi et al., 2003).

Table 3
 Unstandardized Estimates of Indirect Effects: Task-Avoidant Behavior as a Mediator ($N = 586$)

Indirect effect	Estimate (SE)
From externalizing problems via task-avoidant behavior to academic performance	
Externalizing problems (T1) → Task avoidance (T2) → Academic performance (T3)	-0.026 (0.006)***
Externalizing problems (T2) → Task avoidance (T3) → Academic performance (T4)	-0.026 (0.006)***
From academic performance via task-avoidant behavior to externalizing problems	
Academic performance (T1) → Task avoidance (T2) → Externalizing problems (T3)	-0.006 (0.001)***
Academic performance (T2) → Task avoidance (T3) → Externalizing problems (T4)	-0.006 (0.001)***

Note. T = time.

*** $p < .001$ (two-tailed testing of significance).

Second, deficits in attention and self-regulation skills, which are typical externalizing problems (Olson et al. 2005; 2011), are likely to make it difficult for the child to stay focused and finish tasks, which hampers his or her learning and academic performance.

As the second goal of the study, we investigated the association between low academic performance and subsequent externalizing problems and the extent to which this association was mediated by high task-avoidant behavior. Again, academic performance was not directly linked with externalizing problems (Hypothesis 2a). Instead, the results provided consistent support for the reversed mediation model (Hypothesis 2b). Children who showed low academic performance in reading and math in Grades 1 and 2 started to avoid learning tasks in Grades 2 and 3, and eventually had higher rates of externalizing problems. These results are in line with the findings of McGee et al. (1986) indicating reciprocal negative linkages between poor reading skills and externalizing problems. Our findings complement this literature results by showing that the effects of academic skills on externalizing problems are indirect, running through task-avoidant behavior. One possible explanation for the findings is that poor academic performance predisposes children to failure expectations that lead to task avoidance and off-task behavior (Aunola et al., 2002; Lepola et al., 2000; Onatsu-Arvilommi & Nurmi, 2000; Pakarinen et al., 2011). These are likely to cause conflicts with both parents and teachers, which, in turn, exacerbate children's oppositional and acting-out behavior (Arnold, 1997).

Prior research is scarce on the potential mechanisms linking externalizing problems and academic achievement. As an exception, a recent study by Zimmermann et al. (2013) showed that externalizing problems and low academic achievement reciprocally affected each other from middle childhood to adolescence and that low self-esteem partially accounted for this association. The present study extends this line of research and provides new insights to our understanding of the processes by which academic performance and externalizing problems are linked over time. The multiwave cross-lagged panel data allowed investigation of indirect effects via task-avoidant behavior and provided a possibility to examine how behavioral regulation, achievement strategies, and academic learning shape each other over time. The findings suggest a cyclical nature of relationships between the studied variables, which is likely to lead to the strengthening of the difficulties over time. Task-avoidant behavior played a key role in the formation of these reciprocal associations. One explanation for the key role of task-avoidant behavior is that it reflects children's motivational and behavioral ways to approach learning tasks: it is an aspect of motivational orientation that is most closely related to learning in the classroom and homework situations (e.g., Morgan et al., 2008). Nevertheless, the dynamics between externalizing problems and academic achievement are also likely to include additional components, for example, quality of instruction and the nature of the relationship between children and their teachers (Liew, Chen, & Hughes, 2010). These developmental dynamics should be examined in future studies in greater detail.

When interpreting the findings of this study, the following limitations should be considered. The first limitation concerns the data that were drawn from a follow-up sample with a special interest in the risk and protective factors affecting academic skill development and motivation. Consequently, children with an early risk for reading problems were overrepresented in the sample. This

limitation was controlled for by testing the significance of the cross-lagged paths separately for the children with and without risk for reading difficulties. The findings showed that the processes linking externalizing problems, task-avoidant behavior, and academic performance were similar irrespective of kindergarten-age risk for reading difficulties. Similarly, the potential gender differences were examined by testing the paths for girls and boys separately. These analyses demonstrated that the statistically significant paths did not vary systematically by gender. Our findings corroborate those of the previous studies showing that the associations of externalizing problems and various academic measures are similar for boys and girls (Burt & Roisman, 2010; Nelson et al., 2004).

Second, the current study was based on correlational data, and caution should be exercised when interpreting the findings. The study does not provide a definite answer to the question of causality. However, the four-wave cross-lagged panel data allow somewhat stronger inferences about the direction of effects than some earlier cross-sectional studies (Arnold, 1997). Finally, the coefficients for the indirect paths linking externalizing problems to academic achievement through task-avoidant behavior were small in magnitude. This limitation should, however, be viewed against the moderate-to-high stability and within-time co-variation between our constructs, which leave relatively little variance unexplained. Moreover, the effect sizes of the indirect links have also been low in magnitude in many previous studies (e.g., Bornstein et al., 2010; van Lier et al., 2012).

Our findings suggest a need for educational interventions concerning externalizing behaviors that interfere with children's ability to focus and persist in tasks (for a meta-analysis on school-based interventions, see Wilson & Lipsey, 2007). When successful, such interventions would help children to develop more adaptive learning strategies and possibly prevent the development of later mental health problems and antisocial behavior that often follow from early disruptive behaviors (Caspi, 2000; Fergusson et al., 2007). The results of the present study suggest that interventions that aim to reduce task-avoidant behavior by improving classroom quality and teacher-student relationships may also prove useful. Such an intervention would have ramifications both for the development of externalizing problems and academic performance. Previous findings indicate that the higher the quality of teacher's instructional support in the classroom, the less children show task-avoidant behavior (Pakarinen et al., 2011). Teachers can help children to face difficult learning situations by giving individualized feedback and providing tasks that are optimally challenging. They can also encourage children's efforts and thereby decrease their anxiety in learning situations that easily leads to task avoidance. In addition, children's motivation is likely to be fostered by classroom goal orientation, which emphasizes mastery, understanding, and improving skills and knowledge rather than demonstrating high ability or competing for grades (for a review, see Meece, Anderman, & Anderman, 2006).

In addition, interventions targeted specifically to children with maladaptive achievement strategies should also be considered. One central mechanism underlying task-avoidant behavior is the child's negative self-concept and failure expectations. Earlier findings indicate that by fostering children's self-efficacy beliefs, for instance, by informing students of their capabilities and progress in learning and providing learning strategies that help them succeed,

it is possible to support children's task persistence and engagement in learning, and eventually increase their academic knowledge and skills (Schunk & Pajares, 2001). Any prevention or intervention model that is aimed at supporting children's school work should target to multiple domains.

In sum, this study represents a unique effort to investigate the interrelationships among externalizing problems, task-avoidant behavior, and academic performance, and how they shape each other in early school years. The strength of the study lies in the prospective longitudinal design with multiple assessments over time that allowed for testing the mediating effect of task-avoidant behavior in the cross-lagged associations between externalizing behavior and academic performance. In future research, investigators should further examine the developmental trajectories of academic performance and externalizing problems across time and how a more varied set of motivational factors may help to explain the formulation of such trajectories.

References

Adams, J. W., Snowling, M. J., Hennessy, S. M., & Kind, P. (1999). Problems of behavior, reading, and arithmetic: Assessments of comorbidity using the Strengths and Difficulties Questionnaire. *British Journal of Educational Psychology*, 69, 571–585. doi:10.1348/000709999157905

Arnold, D. (1997). Co-occurrence of externalizing behavior problems and emergent academic difficulties in young high-risk boys: A preliminary evaluation of patterns and mechanisms. *Journal of Applied Developmental Psychology*, 18, 317–330. doi:10.1016/S0193-3973(97)80003-2

Atkins, M. S., McKay, M. M., Talbott, E., & Arvanitis, P. (1996). *DSM-IV* diagnosis of conduct disorder and oppositional defiant disorder: Implications and guidelines for school mental health teams. *School Psychology Review*, 25, 274–283.

Aunola, K., Leskinen, E., Lerkkanen, M.-K., & Nurmi, J.-E. (2004). Developmental dynamics of math performance from preschool to Grade 2. *Journal of Educational Psychology*, 96, 699–713. doi:10.1037/0022-0663.96.4.699

Aunola, K., Nurmi, J.-E., Niemi, P., Lerkkanen, M.-K., & Rasku-Puttonen, H. (2002). Developmental dynamics of achievement strategies, reading performance, and parental beliefs. *Reading Research Quarterly*, 37, 310–327. doi:10.1598/RRQ.37.3.3

Aunola, K., & Räsänen, P. (2007). *The 3-Minute Basic Arithmetic Test*. Unpublished test material, Department of Psychology, University of Jyväskylä, Jyväskylä, Finland.

Barry, T. D., Lyman, R. D., & Klinger, L. G. (2002). Academic underachievement and attention-deficit/hyperactivity disorder: The negative impact of symptom severity on school performance. *Journal of School Psychology*, 40, 259–283. doi:10.1016/S0022-4405(02)00100-0

Bongers, I. L., Koot, H. M., van der Ende, J., & Verhulst, F. C. (2003). The normative development of child and adolescent problem behavior. *Journal of Abnormal Psychology*, 112, 179–192. doi:10.1037/0021-843X.112.2.179

Bornstein, M. H., Hahn, C.-S., & Haynes, O. M. (2010). Social competence, externalizing, and internalizing behavioral adjustment from early childhood through early adolescence: Developmental cascades [Special issue]. *Development and Psychopathology*, 22, 717–735. doi:10.1017/S0954579410000416

Broidy, L. M., Tremblay, R. E., Brame, B., Fergusson, D., Horwood, J. L., Laird, R., . . . Vitaro, F. (2003). Developmental trajectories of childhood disruptive behaviors and adolescent delinquency: A six-site, cross-national study. *Developmental Psychology*, 39, 222–245. doi:10.1037/0012-1649.39.2.222

Bub, K. L., McCartney, K., & Willett, J. B. (2007). Behavior problem trajectories and first-grade cognitive ability and achievement skills: A latent growth curve analysis. *Journal of Educational Psychology*, 99, 653–670. doi:10.1037/0022-0663.99.3.653

Burt, K. B., & Roisman, G. I. (2010). Competence and psychopathology: Cascade effects in the NICHD Study of Early Child Care and Youth Development. *Development and Psychopathology*, 22, 557–567. doi:10.1017/S0954579410000271

Campbell, S. B., Shaw, D. S., & Gilliom, M. (2000). Early externalizing behavior problems: Toddlers and preschoolers at risk for later maladjustment. *Development and Psychopathology*, 12, 467–488. doi:10.1017/S0954579400003114

Caspi, A. (2000). The child is father of the man: Personality continuities from childhood to adulthood. *Journal of Personality and Social Psychology*, 78, 158–172. doi:10.1037/0022-3514.78.1.158

Chapman, J. W. (1988). Cognitive-motivational characteristics and academic-achievement of learning-disabled children: A longitudinal study. *Journal of Educational Psychology*, 80, 357–365. doi:10.1037/0022-0663.80.3.357

Coie, J. D., & Dodge, K. A. (1988). Multiple sources of data on social behavior and social status in the school: A cross-age comparison. *Child Development*, 59, 815–829. doi:10.2307/1130578

Deater-Deckard, K., Dodge, K., Bates, J., & Pettit, G. (1998). Multiple risk factors in the development of externalizing behavior problems: Group and individual differences. *Development and Psychopathology*, 10, 469–493. doi:10.1017/S0954579498001709

Duncan, G. J., Dowsett, C. J., Claessens, A., Magnuson, K., Huston, A. C., Klebanov, P., . . . Japel, C. (2007). School readiness and later achievement. *Developmental Psychology*, 43, 1428–1446. doi:10.1037/0012-1649.43.6.1428

Dweck, C. S., & Leggett, E. L. (1988). A social cognitive approach to motivation and personality. *Psychological Review*, 95, 256–273. doi:10.1037/0033-295X.95.2.256

Fergusson, D. M., Horwood, L. J., & Ridder, E. M. (2007). Conduct and attentional problems in childhood and adolescence and later substance use, abuse and dependence: Results of a 25-year longitudinal study. *Drug and Alcohol Dependence*, 88(Suppl. 1), S14–S26. doi:10.1016/j.drugalcdep.2006.12.011

Frick, P. J., Kamphaus, R. W., Lahey, B. B., Loeber, R., Christ, M. A. G., Hart, E. L., & Tannenbaum, L. E. (1991). Academic underachievement and the disruptive behavior disorders. *Journal of Consulting and Clinical Psychology*, 59, 289–294. doi:10.1037/0022-006X.59.2.289

Georgiou, G. K., Manolitsis, G., Nurmi, J.-E., & Parrila, R. (2010). Does task-focused versus task-avoidance behavior matter for literacy development in an orthographically consistent language? *Contemporary Educational Psychology*, 35, 1–10. doi:10.1016/j.cedpsych.2009.07.001

Goodman, R. (1997). The Strengths and Difficulties Questionnaire: A research note. *Journal of Child Psychology and Psychiatry*, 38, 581–586. doi:10.1111/j.1469-7610.1997.tb01545.x

Goodman, R., Ford, T., Simmons, H., Gatward, R., & Meltzer, H. (2000). Using the Strengths and Difficulties Questionnaire (SDQ) to screen for child psychiatric disorders in a community sample. *British Journal of Psychiatry*, 177, 534–539. doi:10.1192/bj.p.177.6.534

Grant, K. E., Compas, B. E., Stuhlmacher, A. F., Thurm, A. E., McMahon, S. D., & Halpert, J. A. (2003). Stressors and child and adolescent psychopathology: Moving from markers to mechanisms of risk. *Psychological Bulletin*, 129, 447–466. doi:10.1037/0033-2909.129.3.447

Gresham, F. M., Lane, K. L., MacMillan, D. L., & Bocian, K. M. (1999). Social and academic profiles of externalizing and internalizing groups: Risk factors for emotional and behavioral disorders. *Behavioral Disorders*, 24, 231–245.

Halonen, A., Aunola, K., Ahonen, T., & Nurmi, J.-E. (2006). The role of learning to read in the development of problem behavior: A cross-lagged longitudinal study. *British Journal of Educational Psychology*, 76, 517–534. doi:10.1348/000709905X51590

Heck, R. H., & Thomas, S. L. (2009). *An introduction to multilevel modeling techniques*. New York, NY: Routledge.

Henricsson, L., & Rydell, A. M. (2004). Elementary school children with behavior problems: Teacher-child relations and self-perception. A prospective study. *Merrill-Palmer Quarterly*, 50, 111–138. doi:10.1353/mpq.2004.0012

Hinshaw, S. P. (1992). Externalizing behavior problems and academic underachievement in childhood and adolescence: Causal relationships and underlying mechanisms. *Psychological Bulletin*, 111, 127–155. doi:10.1037/0033-2909.111.1.127

Hinshaw, S. P., & Lee, S. S. (2003). Oppositional defiant and conduct disorder. In E. J. Mash & R. A. Barkley (Eds.), *Child psychopathology* (2nd ed., pp. 144–198). New York, NY: Guilford Press.

Hirvonen, R., Aunola, K., Alatupa, S., Viljaranta, J., & Nurmi, J. E. (2013). The role of temperament in children's affective and behavioral responses in achievement situations. *Learning and Instruction*, 27, 21–30. doi:10.1016/j.learninstruc.2013.02.005

Hirvonen, R., Tolvanen, A., Aunola, K., & Nurmi, J. E. (2012). The developmental dynamics of task-avoidant behavior and math performance in kindergarten and elementary school. *Learning and Individual Differences*, 22, 715–723. doi:10.1016/j.lindif.2012.05.014

Hox, J. J., & Maas, C. J. M. (2002). Sample sizes for multilevel modeling. In J. Blasius, J. Hox, E. de Leeuw, & P. Schmidt (Eds.), *Social science methodology in the new millennium: Proceedings of the Fifth International Conference on Logic and Methodology* (2nd expanded ed., pp. 1–19). Opladen, Germany: Leske + Budrich Verlag.

Hughes, J. N., Luo, W., Kwok, O., & Loyd, L. K. (2008). Teacher-student support, effortful engagement, and achievement: A 3-year longitudinal study. *Journal of Educational Psychology*, 100, 1–14. doi:10.1037/0022-0663.100.1.1

Jones, E. E., & Berglas, S. S. (1978). Control of attributions about the self through self-handicapping: The appeal of alcohol and the rate of underachievement. *Personality and Social Psychology Bulletin*, 4, 200–206. doi:10.1177/014616727800400205

Ladd, G. W., Birch, S. H., & Buchs, E. S. (1999). Children's social and scholastic lives in kindergarten: Related spheres of influence? *Child Development*, 70, 1373–1400. doi:10.1111/1467-8624.00101

Laird, R. D., Jordan, K. Y., Dodge, K. A., Pettit, G. S., & Bates, J. E. (2001). Peer rejection in childhood, involvement with antisocial peers in early adolescence, and the development of externalizing behavior problems. *Development and Psychopathology*, 13, 337–354. doi:10.1017/S0954579401002085

Leadbeater, B. J., Kuperminc, G. P., Blatt, S. J., & Herzog, D. (1999). A multivariate model of gender differences in adolescent's internalizing and externalizing problems. *Developmental Psychology*, 35, 1268–1282. doi:10.1037/0012-1649.35.5.1268

Lepola, J., Salonen, P., & Vauras, M. (2000). The development of motivational orientations as a function of divergent reading careers from pre-school to the second grade. *Learning and Instruction*, 10, 153–177. doi:10.1016/S0959-4752(99)00024-9

Lerkkanen, M.-K., Ahonen, T., & Poikkeus, A.-M. (2011). The development of reading skills and motivation and identification of risk at school entry. In M. Veisson, E. Hujala, P. K. Smith, M. Waniganayake, & E. Kikas (Eds.), *Global perspectives in early childhood education: Diversity, challenges and possibilities* [Baltische Studien zur Erziehungs- und Sozialwissenschaft; Baltic Studies in Education and Social Science series]. (pp. 237–258). Frankfurt am Main, Germany: Lang.

Lerkkanen, M.-K., Niemi, P., Poikkeus, A.-M., Poskiparta, E., Siekkinen, M., & Nurmi, J.-E. (2006). *Alkuportaat* [The First Steps Study]. Unpublished data, Department of Psychology and Department of Teacher Education, University of Jyväskylä, Jyväskylä, Finland.

Lerkkanen, M.-K., Rasku-Putonen, H., Aunola, K., & Nurmi, J.-E. (2004). Predicting reading performance during the first and the second year of primary school. *British Educational Research Journal*, 30, 67–92. doi:10.1080/01411920310001629974

Liew, J., Chen, Q., & Hughes, J. (2010). Child effortful control, teacher-student relationships, and achievement in academically at-risk children: Additive and interactive effects. *Early Childhood Research Quarterly*, 25, 51–64. doi:10.1016/j.ecresq.2009.07.005

Lindeman, J. (1998). *ALLU—Ala-asteen lukutesti* [Reading test for primary school]. Turku, Finland: University of Turku.

MacKinnon, D. P., & Fairchild, A. J. (2009). Current directions in mediation analysis. *Current Directions in Psychological Science*, 18, 16–20. doi:10.1111/j.1467-8721.2009.01598.x

Mägi, K., Häidkind, P., & Kikas, E. (2010). Performance-approach goals, task-avoidant behavior, and conceptual knowledge as predictors of first graders' school performance. *Educational Psychology*, 30, 89–106. doi:10.1080/01443410903421323

Mägi, K., Lerkkanen, M.-K., Poikkeus, A.-M., Rasku-Putonen, H., & Kikas, E. (2010). Relations between achievement goal orientations and math achievement in primary grades: A follow-up study. *Scandinavian Journal of Educational Research*, 54, 295–312. doi:10.1080/00313831003764545

Masten, A. S., Roisman, G. K., Long, J. D., Burt, K. B., Obradovic, J., Riley, J. R., . . . Tellegen, A. (2005). Developmental cascades: Linking academic achievement and externalizing and internalizing symptoms over 20 years. *Developmental Psychology*, 41, 733–746. doi:10.1037/0012-1649.41.5.733

McConaughy, S. H., Volpe, R. J., Antshel, K. M., Gordon, M., & Eiraldi, R. B. (2011). Academic and social impairments of elementary school children with attention-deficit/hyperactivity disorder. *School Psychology Review*, 40, 200–225.

McGee, R., Williams, S., Share, D. L., Anderson, J., & Silva, P. A. (1986). The relationship between specific reading retardation, general reading backwardness, and behavioral problems in a large sample of Dunedin boys: A longitudinal study from five to eleven years. *Journal of Child Psychology and Psychiatry*, 27, 597–610. doi:10.1111/j.1469-7610.1986.tb00185.x

McLeod, J. D., & Kaiser, K. (2004). Childhood emotional and behavioral problems and educational attainment. *American Sociological Review*, 69, 636–658. doi:10.1177/000312240406900502

McMahon, R. J. (1994). Diagnosis, assessment, and treatment of externalizing problems in children: The role of longitudinal data. *Journal of Consulting and Clinical Psychology*, 62, 901–917. doi:10.1037/0022-006X.62.5.901

Meece, J. L., Anderman, E. M., & Anderman, L. H. (2006). Classroom goal structure, student motivation, and academic achievement. *Annual Review of Psychology*, 57, 487–503. doi:10.1146/annurev.psych.56.091103.070258

Midgley, C., & Urdan, T. C. (1995). Predictors of middle school students' use of self-handicapping strategies. *Journal of Early Adolescence*, 15, 389–411. doi:10.1177/0272431695015004001

Miles, S. B., & Stipek, D. (2006). Contemporaneous and longitudinal associations between social behavior and literacy achievement in a sample of low-income elementary school children. *Child Development*, 77, 103–117. doi:10.1111/j.1467-8624.2006.00859.x

Miller, S. M. (1987). Monitoring and blunting: Validation of a questionnaire to assess styles of information seeking under threat. *Journal of Personality and Social Psychology*, 52, 345–353. doi:10.1037/0022-3514.52.2.345

Miller-Lewis, L. R., Baghurst, P. A., Sawyer, M. G., Prior, M. R., Clark, J. J., Arney, F. M., & Carbone, J. A. (2006). Early childhood externalizing problems: Child, parenting, and family-related predictors over time. *Journal of Abnormal Child Psychology*, 34, 886–901. doi:10.1007/s10802-006-9071-6

Morgan, P. L., Farkas, G., Tufis, P. A., & Sperling, R. A. (2008). Are reading and behavior problems risk factors for each other? *Journal of Learning Disabilities*, 41, 417–436. doi:10.1177/0022219408321123

Murray, C., & Murray, K. M. (2004). Child-level correlates of teacher-student relationships: An examination of demographic characteristics, academic orientations, and behavioral orientations. *Psychology in the Schools*, 41, 751–762. doi:10.1002/pits.20015

Muthén, L., & Muthén, B. O. (1998–2012). *Mplus Version 7.01 & Mplus users' guide*. Los Angeles, CA: Muthén & Muthén.

Nelson, J. R., Benner, G. J., Lane, K., & Smith, B. W. (2004). Academic achievement of K–12 students with emotional and behavioral disorders. *Exceptional Children*, 71, 59–73. doi:10.1177/001440290407100104

Niemi, P., Nurmi, J. E., Lyyra, A. L., Lerkkanen, M. K., Lepola, J., Poskiparta, E., & Poikkeus, A. M. (2011). Task avoidance, number skills, and parental learning difficulties as predictors of poor response to instruction. *Journal of Learning Disabilities*, 44, 459–471. doi:10.1177/0022219411410290

Nurmi, J.-E. (2012). Students' characteristics and teacher-child relationships in instruction: A meta-analysis. *Educational Research Review*, 7, 177–197. doi:10.1016/j.edurev.2012.03.001

Nurmi, J.-E., Aunola, K., Salmela-Aro, K., & Lindroos, M. (2003). The role of success expectation and task-avoidance in academic performance and satisfaction: Three studies on antecedents, consequences, and correlates. *Contemporary Educational Psychology*, 28, 59–90. doi:10.1016/S0361-476X(02)00014-0

Olson, S. L., Sameroff, A. J., Kerr, D. C., Lopez, N. L., & Wellman, H. M. (2005). Developmental foundations of externalizing problems in young children: The role of effortful control. *Development and Psychopathology*, 17, 25–45. doi:10.1017/S09545794050029

Olson, S. L., Tardif, T. Z., Miller, A., Felt, B., Grabell, A. S., Kessler, D., . . . Hirabayashi, H. (2011). Inhibitory control and harsh discipline as predictors of externalizing problems in young children: A comparative study of U.S., Chinese, and Japanese preschoolers. *Journal of Abnormal Child Psychology*, 39, 1163–1175. doi:10.1007/s10802-011-9531-5

Onatsu-Arvilommi, T., & Nurmi, J.-E. (2000). The role of task-avoidant and task focused behaviors in the development of reading and mathematical skills during the first school year: A cross-lagged longitudinal study. *Journal of Educational Psychology*, 92, 478–491. doi:10.1037/0022-0663.92.3.478

Pakarinen, E., Kiuru, N., Lerkkanen, M.-K., Poikkeus, A.-M., Ahonen, T., & Nurmi, J.-E. (2011). Instructional support predicts children's task avoidance in kindergarten. *Early Childhood Research Quarterly*, 26, 376–386. doi:10.1016/j.ecresq.2010.11.003

Risi, S., Gerhardstein, R., & Kistner, J. (2003). Children's classroom peer relationships and subsequent educational outcomes. *Journal of Clinical Child and Adolescent Psychology*, 32, 351–361. doi:10.1207/S15374424JCCP3203_04

Roeser, R., Eccles, J., & Strobel, K. (1998). Linking the study of schooling and mental health: Selected issues and empirical illustrations at the level of the individual. *Educational Psychologist*, 33, 153–176. doi:10.1207/s15326985ep3304_2

Schunk, D., & Pajares, F. (2001). The development of academic self-efficacy. In A. Wigfield & J. Eccles (Eds.), *Development of achievement motivation* (pp. 16–31). San Diego, CA: Academic Press.

Sideridis, G. (2003). On the origins of helpless behavior of students with learning disabilities: Avoidance motivation? *International Journal of Educational Research*, 39, 497–517. doi:10.1016/j.ijer.2004.06.011

Spira, E. G., & Fischel, J. E. (2005). The impact of preschool inattention, hyperactivity, and impulsivity on social and academic development: A review. *Journal of Child Psychology and Psychiatry*, 46, 755–773. doi:10.1111/j.1469-7610.2005.01466.x

Statistics Finland. (2007). *Statistical databases*. Retrieved from http://www.stat.fi/tup/tilastotietokannat/index_en.html

Stipek, D., & Miles, S. (2008). Effects of aggression on achievement: Does conflict with the teacher make it worse? *Child Development*, 79, 1721–1735. doi:10.1111/j.1467-8624.2008.01221.x

Sutherland, K. S., Lewis-Palmer, T., Stichter, J., & Morgan, P. L. (2008). Examining the influence of teacher behavior and classroom context on the behavioral and academic outcomes for students with emotional or behavioral disorders. *Journal of Special Education*, 41, 223–233. doi:10.1177/0022466907310372

van Lier, P. A. C., Vitaro, F., Barker, E. D., Brendgen, M., Tremblay, R. E., & Boivin, M. (2012). Peer victimization, poor academic achievement, and the link between childhood externalizing and internalizing problems. *Child Development*, 83, 1775–1788. doi:10.1111/j.1467-8624.2012.01802.x

Wagner, M., & Cameto, R. (2004). The characteristics, experiences, and outcomes of youth with emotional disturbances [Report from the National Longitudinal Transition Study-2]. *NLTS2 Data Brief*, 3(2). Retrieved from <http://www.ncset.org/publications/>

Wilson, S. J., & Lipsey, M. W. (2007). School-based interventions for aggressive and disruptive behavior: Update of a meta-analysis. *American Journal of Preventive Medicine*, 33, S130–S143. doi:10.1016/j.amepre.2007.04.011

Zhang, X., Koponen, T., Räsänen, P., Aunola, K., Lerkkanen, M.-K., & Nurmi, J.-E. (2014). Linguistic and spatial skills predict early arithmetic development via counting sequence knowledge. *Child Development*, 85, 1091–1107. doi:10.1111/cdev.12173

Zhang, X., Nurmi, J.-E., Kiuru, N., Lerkkanen, M.-K., & Aunola, K. (2011). A teacher-report measure of children's task-avoidant behavior: A validation study of the Behavioral Strategy Rating Scale. *Learning and Individual Differences*, 21, 690–698. doi:10.1016/j.lindif.2011.09.007

Zimmermann, F., Schütte, K., Taskinen, P., & Köller, O. (2013). Reciprocal effects between adolescent externalizing problems and measures of achievement. *Journal of Educational Psychology*, 105, 747–761. doi:10.1037/a0032793

Zuckerman, M., Kieffer, S. C., & Knee, C. R. (1998). Consequences of self-handicapping: Effects on coping, academic performance, and adjustment. *Journal of Personality and Social Psychology*, 74, 1619–1628. doi:10.1037/0022-3514.74.6.1619

Received August 8, 2013

Revision received May 18, 2014

Accepted May 21, 2014 ■

The Big-Fish-Little-Pond Effect: Generalizability of Social Comparison Processes Over Two Age Cohorts From Western, Asian, and Middle Eastern Islamic Countries

Herbert W. Marsh
Australian Catholic University, King Saud University, and
University of Oxford

Alexandre J. S. Morin and Philip Parker
Australian Catholic University

Benjamin Nagengast
University of Tübingen

Adel Salah Abduljabbar
King Saud University

Faisal Abdelfattah
King Saud University

Maher M. Abu-Hilal
Sultan Qaboos University

Extensive support for the seemingly paradoxical negative effects of school- and class-average achievement on academic self-concept (ASC)—the big-fish-little-pond effect (BFLPE)—is based largely on secondary students in Western countries or on cross-cultural Program for International Student Assessment studies. There is little research testing the generalizability of this frame of reference effect based on social comparison theory to primary school students and or to matched samples of primary and secondary students from different countries. Using multigroup–multilevel latent variable models, we found support for developmental and cross-cultural generalizability of the BFLPE based on Trends in International Mathematics and Science Study data; positive effects of individual student achievement and the negative effects of class-average achievement on ASC were significant for each of the 26 groups (nationally representative samples of 4th- and 8th-grade students from 13 diverse countries; 117,321 students from 6,499 classes).

Keywords: big-fish-little-pond effect, Trends in International Mathematics and Science Study, frame of reference effects, contextual effects, doubly-latent multilevel models

Supplemental materials: <http://dx.doi.org/10.1037/a0037485.supp>

Self-concept, dating back to at least the seminal work by William James (1890/1963), is one of the oldest and most important constructs in the social sciences. Today positive self-beliefs are also at the heart of a positive revolution sweeping psychology, which emphasizes focusing on how healthy, normal, and exceptional individuals can get

This article was published Online First September 15, 2014.

Herbert W. Marsh, Institute for Positive Psychology and Education, Australian Catholic University; Department of Education, King Saud University; and Department of Education, University of Oxford; Adel Salah Abduljabbar, Department of Education, King Saud University; Alexandre J. S. Morin and Philip Parker, Institute for Positive Psychology and Education, Australian Catholic University; Faisal Abdelfattah, Department of Education, King Saud University; Benjamin Nagengast, Center for Educational Science and Psychology, University of Tübingen; Maher M. Abu-Hilal, Department of Psychology, Sultan Qaboos University.

This article was made possible in part by a grant from the Australian Research Council (DP130102713). We would like to thank Tihomir Asparouhov, Matthias Von Davier, Anna Preuschoff, and Michael Martin for helpful comments at earlier stages of this research.

Correspondence concerning this article should be addressed to Herbert W. Marsh, Institute for Positive Psychology and Education, Australian Catholic University, Locked Bag 2002, Strathfield NSW 2135, Australia. E-mail: herb.marsh@acu.edu.au

the most from life (e.g., Bandura, 2006; Bruner, 1996; Diener, 2000; Marsh & Craven, 2006; Seligman & Csikszentmihalyi, 2000). Thus, self-concept enhancement is now a major goal in many fields including education, child development, health, sport/exercise sciences, social services, and management (Marsh, 2007). Self-concept is also an important mediating factor that facilitates the attainment of other desirable outcomes. Particularly in education settings, a positive academic self-concept (ASC) is both a highly desirable goal and a means of facilitating subsequent learning and other academic accomplishments. Our study is based on the 2007 Trends in International Mathematics and Science Study (TIMSS), with nationally representative samples from different countries and age cohorts, to provide tests of the developmental and cross-sectional generalizability of strong theoretical models of self-concept using new and evolving statistical methodology.

Big-Fish-Little-Pond Effect (BFLPE): The Theoretical and Substantive Focus

BFLPE Theoretical Models

Self-concept theory emphasizes that perceptions of the self cannot be adequately understood if frames of reference are ignored

(Marsh, 2007). The same objective characteristics and accomplishments can lead to disparate self-concepts, depending on the frames of reference or standards of comparison that individuals use to evaluate themselves, and these self-beliefs have important implications for future choices, performance, and behaviors. From the time of William James (1890/1983), psychologists have recognized that objective accomplishments are evaluated in relation to frames of reference. Thus, James indicated, "we have the paradox of a man shamed to death because he is only the second pugilist or the second oarsman in the world" (p. 310). Marsh (1984; see also Marsh & Parker, 1984; Marsh, Seaton, et al., 2008) proposed the BFLPE to encapsulate frame of reference effects that are based on an integration of theoretical models and empirical research from diverse disciplines (e.g., relative deprivation theory, social comparison theory, psychophysical judgment, social judgment; see supplemental materials).

In the BFLPE model, students are hypothesized to compare their abilities with the abilities of their classmates and use this social comparison impression as one basis for forming their own self-concept (see Figure 1); individual ability is positively related to ASC (the brighter I am, the higher my ASC), but that class- and school-average ability have a negative effect on ASC (the brighter my classmates, the lower my ASC). Hence, ASC depends not only on a student's academic accomplishments but also on those of the student's classmates. Consistent with theoretical predictions and an increasing emphasis on the multidimensionality of self-concept, the BFLPE in academic settings is specific to ASC; class- and school-average ability achievement has little positive or negative effect on nonacademic components of self-concept or on global self-esteem (e.g., Marsh, 1987; Marsh, Chesson, Craven, & Roche, 1995; Marsh & Parker, 1984; for a review, see Marsh, Seaton, et al., 2008).

Diener and Fujita (1997, p. 350) reviewed BFLPE research in relation to the broader social comparison theory and concluded that BFLPE research provided the clearest support for predictions based on social comparison theory in an imposed social comparison paradigm. The reason for this, they surmised, was that the frame of reference, based on classmates within the same class or school, is more clearly defined in BFLPE research than in most other research settings. The importance of the class or school setting is that the relevance of the social comparisons in class or school settings is much more ecologically valid than manipulations

in typical social psychology experiments involving introductory psychology students in contrived settings. Indeed, they argue that except for opting out altogether, it is difficult for students to avoid the relevance of achievement as a reference point within a class or school setting or the social comparisons provided by the academic accomplishments of their classmates (see also Marsh, 2007). Seaton, Marsh, et al. (2008) provided a theoretical rationale for how the BFLPE fits with the broader social comparison research literature, contrasting results for the imposed social comparisons and social comparison when students can freely choose their comparison targets. In support of the direct role of social comparison for the BFLPE, Huguet et al. (2009) demonstrated the BFLPE was largely eliminated after controlling pure measures of social comparison.

Extensive support for the BFLPE generalizes over student groups, subject domains, ASC instruments, and cultures (see reviews by Marsh, Seaton, et al., 2008; Seaton & Marsh, 2013). However, most BFLPE research has been based on high school students from Western developed countries (e.g., Australia, United States, Germany, Israel, France, the Netherlands, and the United Kingdom), as well as in Asian countries such as Hong Kong and Singapore.

Cross-Cultural Support for the BFLPE

In cross-cultural research there are two main orientations, one that focuses on tests of *a priori* hypotheses of cross-cultural differences and one that tests the replicability of existing theories in other cultures and seeks universal, panhuman theories (e.g., Parker et al., 2012; Segall, Lonner, & Berry, 1998, p. 1102). However, strong cross-cultural studies need to compare the results from at least two—and preferably many—countries based on comparable samples and the same measures; otherwise apparent cross-cultural differences are confounded with potential differences in the composition of samples and, perhaps, the appropriateness of materials. Addressing these challenges, there is strong support for the cross-cultural generalizability of the BFLPE for high school students, based on successive data collections of the Organisation for Economic Co-operation and Development Program for International Student Assessment (PISA) data: Marsh and Hau (2003) used the PISA 2000 data based on 103,558 fifteen-year-old students from 26 predominantly industrialized Western countries; Seaton, Marsh, and Craven (2009, 2010) used PISA 2003 (265,180 students, 10,221 schools, 41 countries), which included more collectivist and developing economies than PISA 2000; Nagengast and Marsh (2012) used the PISA 2006 database in the largest cross-cultural study of the BFLPE undertaken to date, and significantly extended the previous PISA studies. In summarizing the three BFLPE-PISA studies, Nagengast and Marsh reported that the effect of school-average achievement was negative in all but one of the 123 samples across the three studies, and significantly so in 114 samples. The average effect size across all 123 samples is $-.223$ (see detailed summary of this previous research in the supplemental materials, Table S3).

The overarching cross-cultural focus of our study was to evaluate the generalizability of the BFLPE in Middle Eastern Islamic countries (where there has been almost no research) to that found in Western countries (that is the basis of most research) and Asian countries (that has been the basis of some research). Although we note why these comparisons are potentially interesting, it was our

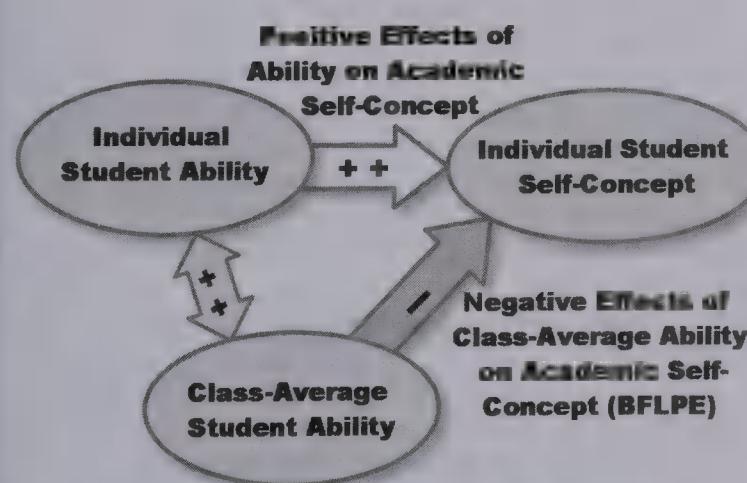


Figure 1. Conceptual model of the big-fish-little-pond effect (BFLPE).

expectation that there would be reasonable support for the generalizability of the results. Indeed, Seaton et al. (2009) claimed support for the universality of the BFLPE as a panhuman theory based on PISA data. However, Schwartz and Bilsky (1990), as well as many others, observed, "Theories that aspire to universality . . . must be tested in numerous, culturally diverse samples" (p. 878). In this respect, one purpose of our study is to greatly expand the scope of tests of the BFLPE's generalizability beyond the set of PISA studies that have been the primary basis of cross-cultural tests of its universality.

PISA versus TIMSS. Although each of the successive PISA studies included a larger and more diverse sample of countries, there were important limitations that are the focus of the present investigation. For example, in their monograph on concerns related to PISA, Hopmann, Brinek, and Retzl (2007; see also Ertl, 2006) summarized a range of substantive, methodological, and policy-related concerns. These included the inappropriateness of the PISA model in respect of what is actually taught in many school systems, technical issues related to translation and scaling, problems with the sampling design, PISA's focus on literacy in testing mathematics, issues in relation to gender, and the league table ranking of countries based on PISA results. Potential concerns such as these dictate that cross-cultural BFLPE research based almost exclusively on PISA data should be cross-validated with data from different sources. Hence, it is surprising that there is apparently no cross-cultural BFLPE research based on the TIMSS data.

Systematic comparisons of results based on the TIMSS and PISA studies (e.g., American Institutes for Research, 2005; Hutchison & Schagen, 2007; National Center for Education Statistics, 2008; Neidorf, Binkley, Gattis, & Nohara, 2006; Wu, 2009) emphasize many similarities but also important differences between achievement tests in the two databases. In particular, PISA focuses more on the application of knowledge to "real life" problems, while TIMSS focuses on achievement more closely linked to school curriculums. Wu (2009) reported that these differences in item content explain in part why Western countries tended to perform better on PISA than TIMSS, while Eastern European and Asian countries tended to perform better on TIMSS than PISA.

A key distinction between TIMSS and PISA that is of particular relevance to the BFLPE lies in the differences in the way data have been collected. PISA samples schools, rather than classrooms, and then tests a random sample of 15-year-olds from each school, so that participants within the same school typically come from two-, three-, or four-year cohorts. Even at the school level, this sampling design complicates interpretation of frame-of-reference effects based on school-average achievement, which typically does not correspond to the achievement levels of students in any of the different year cohorts actually considered. In contrast, although TIMSS also samples schools, it measures all students from selected intact classrooms. However, although TIMSS is nationally representative of each country in relation to classes, there is typically only a single class selected from each school and this class may or may not be representative of the school as a whole. Hence, the appropriate unit of analysis with TIMSS is the classroom rather than the school.

Although the focus of both TIMSS and PISA has been on achievement scores, both databases include a range of psychosocial variables, including ASC responses that are the focus of the present investigation. For both PISA and TIMSS, considerable

effort has been made to make the achievement scores comparable from one data collection to the next, although the rationale for testing achievement differs substantially in PISA and TIMSS. Whereas there has also been reasonable consistency in the items used to infer mathematics self-concept in the different TIMSS data collections (see discussion by Marsh et al., 2013; see also Method section for wording of TIMSS math self-concept items used here), this has not been the case for PISA. First of all, PISA typically only includes math self-concept responses when mathematics is the focus of the PISA data collection. Furthermore, the number and wording of items used in PISA to assess self-concept in the focal domain (math, science, or reading) varies substantially from one data collection to the next (for wording of PISA items in different data collections, see PISA website <http://www.oecd.org/pisa/aboutpisa/>). Finally, on the basis of these comparisons, it is also obvious that the items used to assess self-concept are clearly quite different across the TIMSS and PISA studies.

Local dominance effect. According to the local dominance effect (Zell & Aliche, 2009; see also Liem, Marsh, Martin, McInerney, & Yeung, 2013), the frame of reference—school versus classroom—is a potentially important consideration for BFLPE research. Zell and Aliche (2009; see also Aliche, Zell, & Bloom, 2010) provided support for the BFLPE by experimentally manipulating the frame of reference in relation to feedback given to participants about how their performances compared to others. When they pitted "local" against more "general" comparison standards, participants consistently used the most local comparison information available to them, even when they were told that the local comparison was not representative of the broader population and were provided with more appropriate normative comparison data. Hence, the class-average achievement based on the TIMSS data constitutes a more proximally relevant frame of reference than the school-average achievement based on PISA data, which is likely to be more locally dominant. In this respect, it is important to note that our study is apparently the first cross-cultural BFLPE study to be based on the classroom as the unit of analysis, rather than the school.

Developmental Support for the Generalizability of the BFLPE

For many developmental, educational, and psychological researchers, self-concepts are a "cornerstone of both social and emotional development" (Kagen, Moore, & Bredekamp, 1995, p. 18; see also Davis-Kean & Sandler, 2001; Marsh, Ellis, & Craven, 2002); self-concepts develop early in childhood and, once established, they are enduring (e.g., Eder & Mangelsdorf, 1997). The development of self-concept is therefore emphasized in many early childhood programs (e.g., Fantuzzo, McDermott, Manz, Hampton, & Burdick, 1996).

Hattie (1992; Hattie & Marsh, 1996; see also Eccles, Wigfield, Harold, & Blumenfeld, 1993; Harter, 1999, 2006, 2012; Marsh, Craven, & Debus, 1998) reviewed theoretical and empirical support for stages of growth in the development of self-concept, arguing against the notion of fixed stages that all persons must pass through. Instead, he posited seven parallel developments that are relevant to self-concept formation: (a) children distinguish self and others, (b) children distinguish self and the environment, (c) changes in major reference groups lead to changes in expectations,

(d) attributions are made to salient personal and social or external sources, (e) cognitive processing capacities develop, (f) children develop particular cultural values, and (g) children develop strategies for confirmation and disconfirmation of self-referent information. Thus, with age and development, young children increasingly integrate information from their immediate environment into their self-concept formation. This is particularly relevant to the present investigation, emphasizing the integration of external frames of reference and social comparison into self-concept formation.

Indeed, many authors (Chapman & Tunmer, 1995; Eccles et al., 1993; Harter, 1999; Marsh, 1989; Marsh & Craven, 1997; Skaalvik & Hagtvet, 1990; Wigfield & Eccles, 1992; Wigfield et al., 1997) have offered a developmental perspective on the relation between ASC and academic achievement. For example, Marsh (1989, 1990; Marsh et al., 1998) proposed that the self-concepts of very young children are very positive and are not highly correlated with external indicators (e.g., skills, accomplishments, achievement, self-concepts inferred by significant others) but that with increasing life experience, children learn their relative strengths and weaknesses, so that specific self-concept domains become more differentiated and more highly correlated with external indicators. Marsh et al. (1998) showed that reliability, stability, and factor structure of self-concept scales improve with age (children 5–8 years of age). In addition, consistent with the proposal that children's self-perceptions become more realistic with age, self-ratings of older children were more correlated with inferred self-concept ratings by their teachers.

In summary, there is good developmental theory for the prediction that with age and development ASC becomes more closely related to external criteria, including academic achievement and perceptions of others. From this, it is reasonable to speculate that the BFLPE would also become stronger with age and development, but there is little or no empirical evidence against which to evaluate this supposition. Testing this generalizability of the BFLPE over primary and secondary students is the major focus of our study.

TIMSS 2007: Background to the Present Investigation

An important aspect of TIMSS is collection of data from two age cohorts (corresponding to fourth and eighth grades), providing a unique developmental perspective on cross-cultural studies of the BFLPE. Included in the present investigation are fourth- and eighth-grade classes (see Table 1 for more detail) in six Western countries (Australia, England, Italy, Norway, Scotland, and United States), four Asian countries (Hong Kong, Japan, Singapore, and Taiwan), and three Middle Eastern Islamic countries (Iran, Kuwait, and Tunisia) where both mathematics and science were taught as an integrated subject, and where data were available for both fourth- and eighth-grade cohorts.

In line with the substantial body of BFLPE, we predict that there will be good support for the developmental generalizability of the BFLPE across the two matched age cohorts, and for the cross-cultural generalizability of the BFLPE across the 13 countries. In keeping with developmental models of self-concept (and limited empirical support), we posit that relations between ASC and achievement will be significant for all 26 (2

Table 1

Reliability of the Trends in International Mathematics and Science Study Math and Science Motivation Constructs Used in This Study

Country	Cohort	Number of			Reliability
		Students	Classes	Schools	
Western countries					
Australia	4	4,108	316	228	.747
	8	4,119	238	227	.809
England	4	4,316	233	142	.753
	8	4,046	238	136	.795
Italy	4	4,470	323	169	.687
	8	4,408	287	169	.841
Norway	4	4,108	290	144	.677
	8	4,748	264	138	.805
Scotland	4	3,929	252	138	.723
	8	4,213	244	128	.770
United States	4	7,896	515	256	.763
	8	7,636	510	238	.838
Total	4	28,827	1,929	1,077	.725
	8	29,170	1,781	1,036	.810
Asian countries					
Taiwan	4	4,131	174	149	.735
	8	4,046	153	149	.838
Hong Kong	4	3,791	147	125	.717
	8	3,527	120	119	.803
Japan	4	4,487	189	147	.762
	8	5,625	169	145	.777
Singapore	4	5,041	354	176	.757
	8	4,754	326	163	.825
Total	4	17,450	864	597	.743
	8	17,952	768	576	.811
Middle Eastern Islamic countries					
Iran	4	3,833	224	223	.734
	8	3,981	208	207	.744
Kuwait	4	3,803	181	149	.351
	8	4,091	158	157	.589
Tunisia	4	4,134	217	149	.450
	8	4,080	169	149	.729
Total	4	11,770	622	521	.512
	8	12,152	535	513	.687
All countries					
Grade	4	58,047	3,415	2,195	.681
	8	59,274	3,084	2,125	.781
Total		117,321	6,499	4,320	.731

Note. Reliabilities are expressed as Cronbach's alpha estimates.

age cohorts \times 13 countries) groups but will be stronger for the older age cohort. Of central importance, we predict that the negative effect of class-average achievement on ASC—the BFLPE—will also be significant across all 26 groups. However, we further surmise that the BFLPE will be stronger for the older cohort, in that developmental models posit that social comparison and normative processes in the formation of self-concept grow stronger over this developmental period; but we recognize that there is limited empirical support for this prediction. We leave as a research question whether there are substantively meaningful interactions between country and age cohort differences; alternatively, the extent to which age cohort effects generalize across countries.

Method

Participants

TIMSS 2007 (Olson, Martin, & Mullis, 2008) assessed the competencies in mathematics for nationally representative samples of students from participating countries (for more details about the processes underlying the development of the TIMSS 2007 instruments, translation of materials, sampling, data collection, scaling, and data analysis, see the TIMSS 2007 technical report by Olson et al., 2008). The basic design is a two-stage cluster design that consists of sampling schools and intact classrooms from the target grade in the school. Included in the present investigation are 117,321 students from 6,499 fourth- and eighth-grade classes representing 13 countries (see Table 1 for more detail). In all countries, the materials were administered near the end of the school year (typically October or November in the Southern Hemisphere and April to June in the Northern Hemisphere). For purposes of convenience and consistency with TIMSS 2008, we refer to the fourth-grade cohort (typically 9–11 years of age with 4 years of formal schooling) as primary school children and the eighth-grade cohort (typically 13–15 years of age with 8 years of formal schooling) as secondary school adolescents, but realize that this terminology is not completely consistent across all countries and school systems.

TIMSS (Olson et al., 2008) used item response theory to scale student achievement scores based on a mixture of constructed response and multiple-choice items representing algebra, data and chance, number, and geometry for eighth-grade students and number, geometric shapes and measures, and data display for fourth-grade students. Students in both age cohorts responded to items designed to measure math self-concept (MSC) on the same classic Likert (agree–disagree) response scale; two of the self-concept items had the same wording in the two age cohorts, but there were minor wording changes for the other two self-concept items (see supplemental materials): “I usually do well in math”; “Math is harder for me than for many of my classmates”; “I am just not good at science”; “I learn things quickly in math/science.”

Data Analysis

All analyses, conducted with Mplus 7.0 (Muthén & Muthén, 2013), consisted of multilevel confirmatory factor analyses (CFAs) and structural equation models (SEMs) based on the Mplus robust maximum likelihood estimator, with standard errors and tests of fit that were robust in relation to nonnormality of observations and the use of Likert responses where there were at least four or more response categories, particularly where nonnormality was not excessive (e.g., Beauducel & Herzberg, 2006; DiStefano, 2002; Dolan, 1994; Muthén & Kaplan, 1985). Maximum likelihood estimation is also robust to the nonindependence of the observations when used in conjunction with a design-based correction (Mplus’s complex design option; Muthén & Muthén, 2013). All analyses were based on TIMSS’s HOUWGT weighting variable that incorporates three components related to sampling of the school, class, and student, respectively, and three associated with nonparticipation at the level of the school, class, and student. For present purposes, the 26 (13 countries \times 2 age cohorts) groups were treated as grouping variables that were the basis of the

multigroup analyses, whereas the class and school identification variable was treated as a clustering variable to control for the clustered sample (using the complex design option and robust maximum likelihood options in Mplus; Muthén & Muthén, 2013).

In the TIMSS 2007 database, the achievement tests for each student are reported as five plausible values (Olson et al., 2008)—numbers drawn randomly from the distribution of scores that could be reasonably assigned to each student. Although the amount of missing data was small (an average of less than 2% per item), we used full-information maximum likelihood estimation to control for missing data, noting that this had been done separately for each of the five data sets based on different plausible values and then combined using the Rubin (1987; Schafer, 1997) strategy to obtain unbiased parameter estimates, standard errors, and goodness-of fit statistics.

Comparison of results across different countries, and age cohorts, requires strong assumptions about the invariance of the factor structure. If the underlying factors differ fundamentally in different groups, then there is no basis for interpreting observed differences (the “apples and oranges” problem; see Millsap, 2011). Here we initially consider invariance across the 26 (13 countries \times 2 age cohorts) groups. In applied SEM research—particularly for large sample sizes as in TIMSS—there is a predominant focus on indices that are sample size independent (e.g., Hu & Bentler, 1999; Marsh, Balla, & McDonald, 1988; Marsh, Hau, & Grayson, 2005; Marsh, Hau, & Wen, 2004). The Tucker–Lewis index (TLI) and the comparative fit index (CFI) vary along a 0–1 continuum, and values greater than .90 and .95 typically reflect acceptable and excellent fit to the data, respectively. For the root-mean-square error of approximation (RMSEA), values of less than .05 and .08 reflect a close fit and a minimally acceptable fit to the data, respectively. However, for purposes of model comparison, comparison of the relative fit of models imposing more or fewer invariance constraints, Cheung and Rensvold (2002) and Chen (2007) suggest that if the decrease in fit for the more parsimonious model is less than .01 for incremental fit indices like the CFI, then there is reasonable support for the more parsimonious model. Chen suggests that when the RMSEA increases by less than .015, there is support for the more constrained model. However, these guidelines are based on simulated data studies and practice typically involving only two or a small number of groups, and might not be fully applicable to studies like the present investigation, based on 26 groups. Hence, it is important to emphasize that these are only rough guidelines (Marsh, Hau, & Wen, 2004), and it is recommended that applied researchers use an eclectic approach based on subjective integration of a variety of different indices—including the chi-square, detailed evaluations of parameter estimates in relation to theory, a priori predictions, common sense, and alternative models specifically designed to evaluate goodness of fit in relation to key issues. This is consistent with the approach we used here.

Latent contextual effect models: Substantive-methodological synergy. Only recently has BFLPE research integrated the application of multilevel models (e.g., students nested within classes) with the use of latent variable models (with multiple indicators of latent constructs, the multiple MSC items) and multiple group analyses to compare results across countries (see Lüdtke, Marsh, Robitzsch, & Trautwein; Lüdtke et al., 2008; Marsh, Lüdtke, et al., 2009; Nagengast & Marsh, 2012). In the present application of this evolving statistical approach (see Figure 2), we used manifest aggregation to form the class-average measure of achievement such that

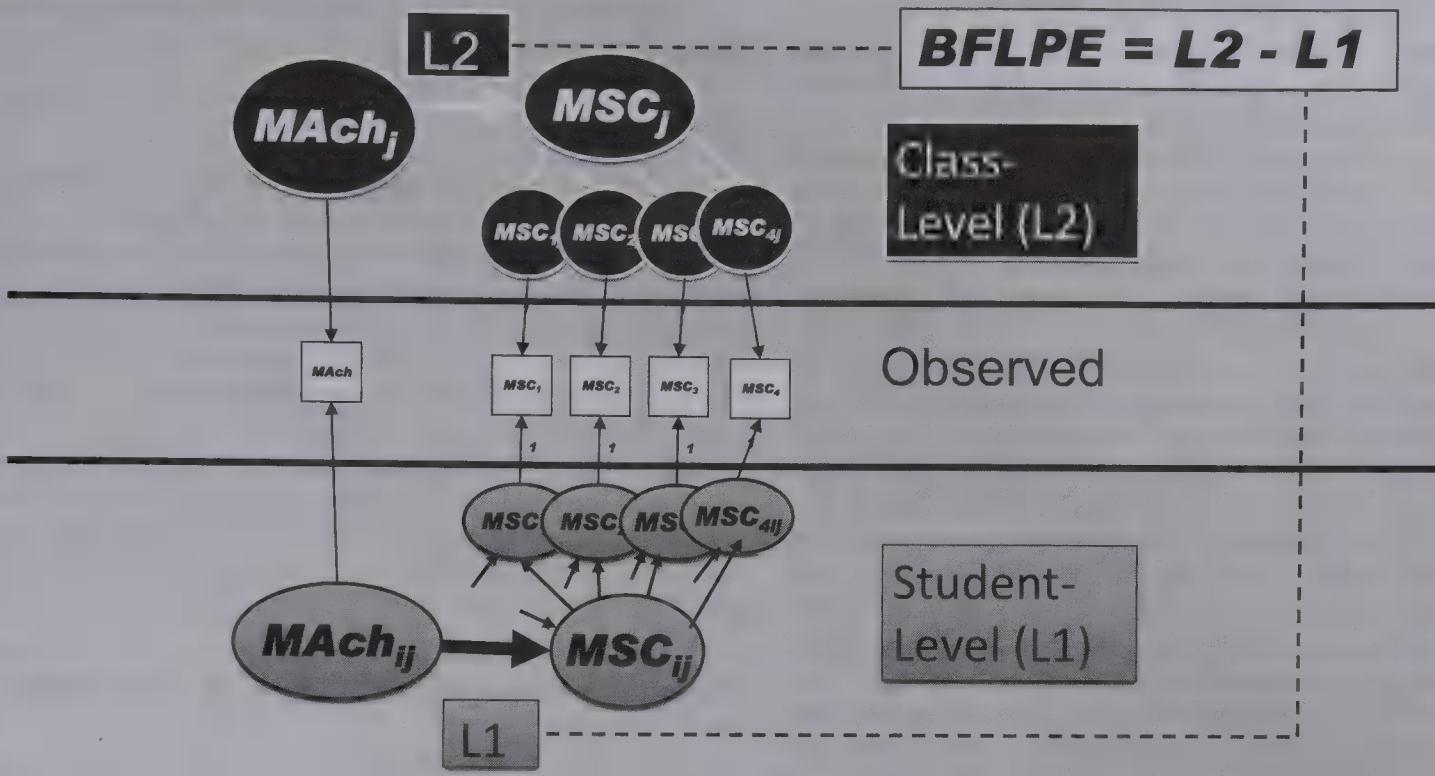


Figure 2. Multilevel depiction of the big-fish-little-pond effect (BFLPE). MAch = math achievement; MSC = math self-concept.

$$\begin{aligned} \text{Self-concept} &= \beta_0 + \beta_1 (\text{achievement}) + r \\ \beta_0 &= \gamma_{00} + \gamma_{01} (\text{mean}_{\text{achievement}}) + \mu_0, \end{aligned} \quad (1)$$

where β_0 is a random intercept and β_1 is the effect of individual student achievement on self-concept; γ_{01} represents the variation in β_0 that is explained by school-average achievement; r and μ_0 are residual terms. For TIMSS data, intact classes were sampled so that the sampling ratio approached 1.0 and so sampling error was minimal. Indeed, the use of latent aggregation (and the use of within-class achievement variation to estimate sampling error) would overcorrect BFLPE estimates (see Marsh, Lüdtke, et al., 2009; although the size of the bias would be small because of the substantial sample sizes).

The contextual effects models were estimated with the reflective aggregation procedure in Mplus (Muthén & Muthén, 2013) that uses implicit group-mean centering of all L1 variables. This implies that the partial regression weights associated with L1 variables reflect L1 effects, while the partial regression weights associated with L2 variables reflect L2 effects that are not controlled for L1 differences (Enders & Tofghi, 2007; Kreft, de Leeuw, & Aiken, 1995). Estimates

of contextual effects, that represent the effect of L2 variables after controlling for L1-differences, can be obtained by subtracting the L1 effect from the L2 effect (Enders & Tofghi, 2007; Kreft et al., 1995):

$$\beta_{\text{context}} = \beta_{\text{L2}} - \beta_{\text{L1}}, \quad (2)$$

where β_{L2} is the L2 effect, β_{L1} is the L1 effect, and β_{context} is the contextual effect (see Figure 2). The standard error for the contextual effect was obtained with the multivariate delta method (see Raykov & Marcoulides, 2004).

In order to facilitate comparisons with previous research, effect sizes (ESs) for the BFLPE (the effect of class-average achievement on MSC after controlling for individual student achievement) were calculated according to the recommendations of Marsh, Lüdtke, et al. (2009; Nagengast & Marsh, 2012) by the following formula:

$$\text{BFLPE ES} = 2 * \beta * \sigma_{\text{pred}} / \sigma_y, \quad (3)$$

where β is the unstandardized regression coefficient, σ_{pred} is the standard deviation of the predictor variable (achievement), and σ_y

is the standard deviation of the outcome variable (self-concept), resulting in an ES metric that is common across countries. This ES is comparable to Cohen's d (Cohen, 1988), reflecting differences based on classes 1 standard deviation above the mean and 1 standard deviation below the mean. For MSC, students in all 26 (13 countries \times 2 cohorts) groups completed the same items, and so it was appropriate to standardize MSC responses in relation to a standard deviation that was common across all 26 groups. However, for math achievement (MAch), the tests were completely different for the two age cohorts, so that we standardized the achievement scores to have $M = 0$ and $SD = 1$ across the 13 countries within each of the two age cohorts.

In the decomposition of group (13 countries \times 2 age cohorts) into variance components and more detailed factorial (analysis-of-variance-like) contrasts, we relied heavily on the flexibility of the "model constraint" function in Mplus. The resulting tests of statistical significance based on these model constraints were based on the delta method (Muthén & Muthén, 2013). Thus for example, we used these constraints to obtain analysis-of-variance-like estimates of the statistical significance and proportion of variation in the relations of MSC with student level (L1) achievement and class-average (L2) achievement that was explained by the 13 countries (and three groups of countries: Western, Asian, Middle Eastern Islamic), two age cohorts (Grade 4 versus Grade 8), and Age Cohort \times Country interactions. These were followed by more specific tests of a priori hypotheses. This evolving methodology—combining the flexibility typically associated with analyses of manifest variables with latent variable models—is apparently a new contribution.

Preliminary Results

In preliminary analyses we estimated the average reliability of the MSC score and evaluated the a priori factor structure for these responses based on Marsh et al. (2013; see supplemental materials for more detail). Due in part to the brevity of the four-item MSC scale, reliability estimates (see Table 1) sometimes reached a desirable standard of .80, but in other cases fell below acceptable values of .70 or even .60. Reliability estimates were systematically higher for the older age cohort (mean $\alpha = .781$) than the younger cohort (mean $\alpha = .681$), and substantially lower in the Middle Eastern Islamic countries than in the Western or Asian countries. These systematic differences in reliability make problematic comparisons based on manifest scale or composite scores, and support the need to consider latent variable models that control for unreliability.

Our a priori factor model (following from Marsh et al., 2013; see supplemental materials) is a simple model in which the four self-concept items are associated with one latent self-concept factor, MAch is a single-item variable (represented by TIMSS's five sets of plausible values that control for unreliability), and there is a negative-item method effect represented by a correlated uniqueness between the two negatively worded self-concept items. This model was supported based on a series of single-level multigroup (using the Mplus complex design to control for clustering of students within classes and schools) tests of invariance over 26 (2 cohorts \times 13 countries) groups. Next we tested multilevel-multigroup CFA models demonstrating the invariance of factor loadings within and across student (L1) and class (L2) levels as

well as the 26 groups. Subsequent results supported the highly constrained model, in which all factor loadings were constrained to be the same across all 26 groups at both the student and class levels ($CFI = .956$, $TLI = .941$, $RMSEA = .054$; see supplemental materials for the Mplus syntax used to test this model), that was the basis of subsequent results.

Results

Support for the BFLPE requires that the effect of individual (L1) achievement on MSC is positive, while the effect of class-average (L2) achievement is negative (see Figure 1). The standardized path coefficients between individual student achievement and MSC are significantly positive in all 26 (13 countries \times 2 cohorts) groups ($M = .592$, $SE = .005$; see Table 2). In contrast, the BFLPEs (ESs for the negative effect of class-average achievement on MSC) are significantly negative in all 26 groups ($M = -.377$, $SE = .012$; see Table 2). These results provide strong support for the generalizability across countries and across age cohorts. We now address substantively important developmental and cross-cultural issues, evaluating how these effects vary as a function of age cohort, country, and their interaction.

Relations Between Student Level (L1) Achievement and Self-Concept

Averaged across all countries and age cohorts, achievement and MSC are positively correlated ($M = .592$, $SE = .005$; see Table 2). Next, we decomposed variance estimates into contrast tests of differences associated with the 13 countries, the two age cohorts, and their interactions; and estimated variance components for each of these differences (sums of squares and variance components in Table 2). Estimates for all 26 groups, the mean estimate for each country, the mean estimate for each cohort, and the means for each of the three country groupings are all significant and positive. The variance components associated with each effect—along with an inspection of the values for each of the 26 (2 cohorts \times 13 countries) groups—provide an indication of the sizes of the effects and how well they generalize over age cohorts and countries.

Cohort effects. The size of the relation between MAch and MSC, averaged across all countries, is substantially larger for the older cohort ($M = .692$, $SE = .008$) than for the younger cohort ($M = .492$, $SE = .006$). However, interpretations of cohort differences are complicated by Cohort \times Country interactions, suggesting that cohort differences vary for different countries. The positive relations are substantially larger for secondary than primary students (i.e., difference scores in Table 2 are significantly positive) in all Western and in all Islamic countries, but not in all the Asian countries. Although the mean estimate across the Asian countries is significantly ($p < .05$) higher for the older cohort, the difference is small and inconsistent across the Asian countries. Only in Singapore is the positive relation larger for secondary than primary students, and in Taiwan the positive relation is significantly larger for primary than for secondary students (cohort differences are not significant in Hong Kong and Japan).

Country differences. Although the relation between self-concept and achievement is significantly positive for both cohorts in each of the 13 countries (see Table 2), estimates are more positive in the Western ($M = .627$, $SE = .008$) and Asian ($M =$

Table 2

Effects of Individual Student Achievement and Class-Average Achievement on Math Self-Concept Broken Down by 13 Countries \times 2 Age Cohorts

Country	Age cohort	Individual achievement	Class-average achievement (BFLPE effect size)
Western countries			
Australia	4	.583 (.021)	-.358 (.042)
	8	.914 (.032)	-.627 (.063)
	Diff	.331 (.038)	-.269 (.075)
	Total	.749 (.019)	-.493 (.038)
England	4	.460 (.018)	-.294 (.048)
	8	.629 (.035)	-.359 (.051)
	Diff	.169 (.039)	-.065 (.069)*
	Total	.544 (.019)	-.327 (.035)
Italy	4	.481 (.019)	-.482 (.041)
	8	.832 (.020)	-.907 (.068)
	Diff	.351 (.028)	-.425 (.079)
	Total	.656 (.014)	-.694 (.039)
Norway	4	.369 (.019)	-.134 (.054)
	8	.937 (.022)	-.527 (.086)
	Diff	.568 (.030)	-.393 (.104)
	Total	.653 (.014)	-.331 (.050)
Scotland	4	.364 (.019)	-.418 (.072)
	8	.563 (.032)	-.282 (.058)
	Diff	.199 (.036)	.137 (.093)*
	Total	.463 (.019)	-.350 (.046)
United States	4	.631 (.018)	-.352 (.038)
	8	.766 (.025)	-.502 (.050)
	Diff	.135 (.033)	-.150 (.065)
	Total	.699 (.014)	-.427 (.030)
Mean Western	4	.481 (.009)	-.340 (.021)
	8	.774 (.013)	-.534 (.026)
	Diff	.292 (.016)	-.194 (.034)
	Total	.627 (.008)	-.437 (.016)
Asian countries			
Hong Kong	4	.609 (.032)	-.441 (.062)
	8	.636 (.034)	-.549 (.051)
	Diff	-.027 (.054)*	-.107 (.085)*
	Total	.623 (.020)	-.495 (.038)
Japan	4	.578 (.019)	-.247 (.078)
	8	.574 (.019)	-.482 (.068)
	Diff	.004 (.027)*	-.236 (.106)
	Total	.576 (.013)	-.364 (.051)
Taiwan	4	.688 (.026)	-.475 (.077)
	8	.581 (.019)	-.180 (.056)
	Diff	-.107 (.028)	.295 (.095)
	Total	.635 (.018)	-.327 (.047)
Singapore	4	.589 (.019)	-.211 (.040)
	8	.828 (.033)	-.585 (.062)
	Diff	.239 (.036)	-.374 (.074)
	Total	.708 (.020)	-.398 (.037)
Mean Asia	4	.616 (.011)	-.343 (.033)
	8	.655 (.016)	-.449 (.031)
	Diff	.039 (.018)	-.106 (.047)
	Total	.635 (.010)	-.396 (.022)
Middle Eastern Islamic countries			
Iran	4	.510 (.022)	-.175 (.044)
	8	.580 (.023)	-.362 (.047)
	Diff	.069 (.031)	-.186 (.064)
	Total	.545 (.016)	-.268 (.032)
Kuwait	4	.236 (.017)	-.089 (.038)
	8	.457 (.023)	-.342 (.065)
	Diff	.221 (.027)	-.252 (.076)
	Total	.347 (.015)	-.216 (.037)

Table 2 (continued)

Country	Age cohort	Individual achievement	Class-average achievement (BFLPE effect size)
Tunisia	4	.300 (.014)	-.117 (.048)
	8	.703 (.026)	-.314 (.102)
	Diff	.403 (.030)	-.197 (.112)*
	Total	.502 (.015)	-.216 (.057)
Mean Islam	4	.349 (.010)	-.127 (.025)
	8	.580 (.014)	-.339 (.043)
	Diff	.231 (.016)	-.212 (.050)
	Total	.464 (.009)	-.233 (.025)
Mean across all 26 (13 countries \times 2 cohorts) groups			
Total	4	.492 (.006)	-.292 (.015)
	8	.692 (.008)	-.463 (.018)
	Diff	.200 (.010)	-.171 (.023)
	Total	.592 (.005)	-.377 (.012)
Sums of squares (SS) and variance components (VC) ^a			
SS cohort		.260 (.026)	.210 (.053)
	VC	.067	.013
SS country		.300 (.033)	.419 (.077)
	VC	.077	.026
SS interaction		.205 (.021)	.226 (.060)
	VC	.053	.014

Note. For each country, the three country groupings, and the total across all 26 (13 countries \times 2 cohorts) groups, results are presented for each age cohort (fourth grade and eighth grade), the difference (diff) between age cohorts, and the total across age cohorts. For each estimate there is a standard error (in parentheses) that can be used to assess statistical significance (i.e., estimates divided by their standard error that are greater than 1.96 are statistically significant at $p < .05$). For individual student achievement, the estimates are the standardized path coefficients, while for the big-fish-little-pond effect (BFLPE) the estimates are effect sizes.

* Effects across the 26 groups were decomposed to assess the main effects of differences due to the 13 countries, the two age cohorts, and their interaction (sums of squares and variance components).

^a Estimates that are not statistically significant in the predicted direction; all other estimates are statistically significant at $p < .05$.

.635, $SE = .010$) than in the Islamic ($M = .464, SE = .009$) countries. However, results for individual countries are not entirely consistent, even within each of the three country classifications, and these differences interact significantly with cohort. For the younger cohort, the relations are substantially more positive in Asian countries ($M = .616, SE = .011$) than in the Western ($M = .481, SE = .009$) and particularly the Islamic ($M = .349, SE = .010$) countries. However, for the older cohort, relations are substantially more positive in the Western countries ($M = .774, SE = .013$) than in Asian ($M = .616, SE = .011$) and particularly in Islamic ($M = .580, SE = .014$) countries. Averaged across cohorts, the estimates are most positive in Australia, Singapore, and the United States, but are also higher in Italy, Norway, Taiwan, Hong Kong, and Japan than in any of the three Islamic countries.

The BFLPE: The Negative Effects of Class-Average Ability on MSC

Averaged across all countries and age cohorts, class-average achievement has a negative effect on MSC (mean BFLPE $ES = -.377, SE = .012$; see Table 2). Although the BFLPE is significantly negative for each of the 26 (2 cohorts \times 13 countries)

groups, its size does vary significantly with cohort, country, and their interaction, as demonstrated by variance components—along with an inspection of the values for each of the 26 groups.

Cohort effects. Because there have been no previous cross-cultural studies of the BFLPE with primary school students, the most important result of our study is that in all 13 countries the BFLPE is statistically significant and negative for the younger cohort, as well as the older cohort (see Table 2). An important contribution of our study is the finding—consistent with predictions—that the BFLPE is significantly larger in the eighth-grade cohort (mean BFLPE ES = $-.463$, $SE = .018$) than in the fourth-grade cohort (mean BFLPE ES = $-.283$, $SE = .015$). Furthermore, the mean BFLPE ESs are significantly more negative for the older cohort in each of the three country groups, and these cohort differences do not differ significantly from each other (West: $-.194$, $SE = .065$; Asian: $-.106$, $SE = .047$; Islamic: $-.212$, $SE = .050$). Again, however, interpretations of cohort differences are complicated by Cohort \times Country interactions. For four countries (England, Scotland, Hong Kong, and Tunisia) the BFLPE did not differ significantly as a function of cohort, while in one country (Taiwan) the BFLPE was significantly more negative for the younger cohort.

Country differences. The BFLPE is significantly negative for both cohorts in all 13 countries, but there are differences between countries (see Table 2). Across the cohorts, the BFLPE is more negative in the Western (mean BFLPE ES = $-.437$, $SE = .016$) and Asian (mean BFLPE ES = $-.396$, $SE = .022$) countries than in Islamic (mean BFLPE ES = $-.233$, $SE = .025$) countries. As already noted, the BFLPE is noticeably smaller in the younger cohort of Islamic countries ($-.127$, $SE = .025$). The BFLPE is particularly large in Italy (mean BFLPE ES = $-.694$, $SE = .039$), but is also very substantial in Hong Kong (mean BFLPE ES = $-.495$, $SE = .038$), Australia (mean BFLPE ES = $-.493$, $SE = .038$), and, to a lesser extent, the United States (mean BFLPE ES = $-.427$, $SE = .030$) and Singapore (mean BFLPE ES = $-.398$, $SE = .037$). The BFLPE was smallest in Tunisia (mean BFLPE ES = $-.216$, $SE = .057$) and Kuwait (mean BFLPE ES = $-.216$, $SE = .037$), particularly for the younger cohort.

Discussion

Substantively and theoretically, the most important result of the present investigation is that the BFLPE—the negative effect of class-average achievement on MSC—is statistically significant and generalizes across both age cohorts in all 13 countries, providing good support for its developmental and cross-national generalizability. This is important because this is the only large-scale cross-cultural study to compare the BFLPE across matched samples of primary and secondary students from a broad array of different countries.

Our study is also the first large-scale cross-cultural study of the BFLPE not based on PISA. Importantly, the consistency of the BFLPEs for both cohorts for the TIMSS data in our study is even stronger than in previous cross-cultural studies based on PISA data. Thus, the average BFLPE ES across 123 samples based on PISA data (59 countries sampled in one or more data collections in PISA 2000, PISA 2003, and PISA 2006) is $-.223$, while the average BFLPE ES across 24 samples (12 countries \times 2 age cohorts) in the present study is $-.377$. Furthermore this general

trend is reasonably consistent across overlapping countries that participated in both PISA and TIMSS. This might seem surprising, in that PISA data is based on somewhat older students—15-year-olds—than even the oldest TIMSS cohort, and our results, consistent with a priori development predictions, show that the BFLPE is more negative for older students ($-.292$ for Year 4, $-.426$ for Year 8). However, these findings are consistent with our a priori predictions based on the local dominance effect when comparing results based on school-average achievement (PISA) and class-average achievement (TIMSS). Nevertheless, there are a number of critical differences between TIMSS and PISA sampling designs that might explain, in part, these differences but also dictate caution in interpretation of the results:

- The nature of the standardized achievement tests is more closely related to the academic curriculum in TIMSS than in PISA, so that the frame of reference based on class-average TIMSS is more closely associated with the achievement results (e.g., school grades, teacher feedback) that are actually experienced by these students.
- The sampling unit for PISA is the whole school, while that of TIMSS is the individual class. Although both are relevant, there is some theoretical and empirical research (see earlier discussion of the local dominance effect; e.g., Liem et al. 2013; Zell & Aliche, 2009) suggesting that the more proximal frame of reference associated with the class is stronger—more locally dominant—than that associated with the whole school, particularly if there is streaming or tracking within schools, so that there are systematic differences in achievement levels for different tracks within schools.
- TIMSS samples intact classes that almost always represent a single-year group, whereas PISA samples 15-year-olds and typically includes two, three, or even four year groups within a given school. Thus, BFLPE interpretations for PISA are complicated in that the school-average ability estimate is an aggregation of test scores across multiple-year groups that do not correspond to any one of the year groups actually considered. Also, school-average ability estimates in PISA studies are typically based on a relatively small proportion of the 15-year-olds in the school, so that at least moderate amounts of sampling error in the school-average estimates are likely; the school-average estimate is a sample mean estimate of the true school-average value if all 15-year-olds in the school are tested. Although recent advances in contextual models used to assess the BFLPE are able to control for sampling error (Marsh et al., 2009; Nagengast & Marsh, 2012), this is typically at the expense of statistical power. In contrast, class-average estimates of achievement based on TIMSS scores are based on intact classes, so that there is little or no sampling error.

In summary, our results demonstrate that the size of the BFLPE is systematically larger for high school students than primary school students, although the BFLPE is clearly evident in both age cohorts. Our findings also suggest, consistent with the local dominance effect, that the BFLPE is substantially more negative for class-average achievement, as in TIMSS data, than for school-average achievement, as in PISA data. However, neither PISA nor TIMSS is ideal for testing this difference in that neither of these international comparisons allows researchers to distinguish properly between the effects of class- and school-average achievement in a more appropriate three-level (L1 = students, L2 = classes, L3 = schools) model. Furthermore, although it is likely that future

research will be able to address this issue with data from individual countries, it seems unlikely that that future research will be able to test the cross-cultural generalizability of these results with data as comprehensive as the PISA and TIMSS data. However, a comprehensive meta-analysis of BFLPE studies (that included PISA and TIMSS studies as well as the large number of studies done in individual countries) would be a useful addition to the literature.

Our study is also apparently the first to specifically compare BFLPE results in a sample of Middle Eastern Islamic countries with those from Asian and Western countries, which have been the basis of most BFLPE research. Indeed, only one of these Islamic countries in our study (Tunisia) had participated in PISA. Although the BFLPE was statistically significant for both age cohorts in all the Islamic countries, it was significantly smaller—particularly in the younger age cohort. In line with earlier research in Arab and Islamic countries (Marsh et al., 2013; see also Abu-Hilal & Bahri, 2000), we also found that relations between L1 achievement and MSC were significantly smaller in the Islamic countries—again, particularly for the younger age cohort. These authors previously speculated that students from these countries do not receive as much evaluative feedback about their achievement as Western and Asian students, and are not socialized in such a way as to critically evaluate their academic skills in relation to classmates. Indeed, consistent with speculations by Abu-Hilal and Bahri (2000) that self-concept formation of ASC and its relation with achievement in Middle Eastern Islamic middle school students was similar to that found in younger students from Western countries, support for the BFLPE for eighth-grade Middle Eastern Islamic students is similar to that found for the fourth-grade cohort in the Western and Asian countries (for further discussion, see Abu-Hilal, 2001; Abu-Hilal & Aal-Hussain, 1997; Abu-Hilal & Bahri, 2000).

Strengths, Limitations, and Directions for Further Research

Important strengths of our study include the use of large, nationally representative samples of primary and secondary school students from culturally diverse countries who were tested with standardized materials under standardized conditions; the integration of CFA, SEM, and multiple-group and multilevel modeling into a single analytic framework; and decomposition of differences in the BFLPEs associated with age cohort, country, and their interaction. In these respects, our study is a strong exemplar of the methodological-substantive synergies that apply evolving statistical methodology to substantively and theoretically important issues that have policy and practice implications. Nevertheless, as is always the case, there are important limitations that may provide the basis of further research.

Reliance on cross-sectional data for only two age cohorts. Reliance on only two cross-sectional age cohorts requires additional caveats in the interpretation of the results from a developmental perspective. For example, the apparent differences as a function of age might also be a function of birth cohort effects, and we were not able to evaluate how consistent the effects of age were for different individuals. Nevertheless, there are also some limitations with longitudinal data (e.g., generalizability of the results to other age cohorts; complications in sampling designs, missing data, representativeness of data within each country, and compa-

rability across countries—particularly in relation to tracking students from primary to secondary schools). Ultimately, the “best” developmental description of how the BFLPE must incorporate findings from both cross-sectional and longitudinal studies, more fully evaluate developmental aspects of the BFLPE, and use a wider range of ages based on a combination of multicohort and longitudinal data.

Assumptions of causality and underlying processes. Support for the BFLPE—and contextual models more generally—is largely based on cross-sectional, correlational studies, so that causal interpretations should be offered tentatively and interpreted cautiously. In particular, the “third variable” problem is always a threat to contextual studies that do not involve random assignment, but Marsh, Hau, and Craven (2004; Marsh, Seaton, et al., 2008) argue that this is an unlikely counterexplanation of BFLPE results, in that most potential third variables (resources, per student expenditures, socioeconomic status, teacher qualifications, etc.) are positively related to class- or school-average achievement, so that controlling for them would typically increase the size of the BFLPE. Fortunately, there is now a growing body of BFLPE research using various combinations of longitudinal, quasi-experimental, and true experimental designs that all support the BFLPE (see Marsh, 2007; Marsh, Seaton, et al., 2008; Nagengast & Marsh, 2012). Quasi-experimental, longitudinal studies (e.g., Marsh, Kong, & Hau, 2000) show that students’ ASC declines when students shift from mixed-ability schools to academically selective schools over time (based on pre- and posttest comparisons) and compared to students matched on academic ability who continue to attend mixed-ability schools. There is support for the BFLPE in studies where achievement is based on tests administered before students began high school (e.g., Marsh et al., 2000). Extended longitudinal studies (Marsh et al., 2000; Marsh, Trautwein, Lüdtke, Baumert, & Köller, 2007) show that the BFLPE grows more negative the longer students attend a selective school and is maintained even 2 and 4 years after graduation from high school. Also, there is good support for the theoretical underpinnings of the BFLPE, as it is largely limited to academic components of self-concept and nearly unrelated to nonacademic components of self-concept and self-esteem (Marsh, 1987; Marsh & Parker, 1984). However, further longitudinal and intervention studies would be useful to bolster the case for mediation of the effects of L1 and L2 achievement on subsequent achievement and educational attainment by ASC.

Also implicit in the BFLPE is the assumption that the direction of causal ordering is from class- or school-average ability to ASC. Although apparently reasonable, this implicit causal ordering cannot easily be tested with cross-sectional data. However, there are also studies in support of the BFLPE based on longitudinal data where the temporal ordering is more clear-cut and for true experimental studies in which class- or school-average achievement is experimentally manipulated.

Policy Implications

Our study greatly extends the generality of the negative effects of attending classes and schools where the average ability level of classmates is high, demonstrating the cross-cultural generalizability to primary school children as well as secondary school adolescents. These results also greatly expand the scope of support for

the universality of the BFLPE as a panhuman theory that has previously been based primarily on PISA data (Seaton et al., 2009). Indeed, our results suggest that the negative effects of school-average achievement based on PISA might substantially underestimate the results for more proximally relevant measures of class-average achievement based on TIMSS. Although theoretically important, these findings are worrisome, as ASC is well known to be an important predictor of academic choice and long-term engagement (Guay, Marsh, & Boivin, 2003; Marsh, 1991; Marsh & Craven, 2006; Marsh & O'Mara, 2010; Marsh & Yeung, 1997). Particularly when so many parents, teachers, and policy analysts uncritically assume that academic selective schools must automatically benefit the students who attend them, it is important to provide an alternative perspective based on strong theory and rigorous research. More generally, BFLPE research provides an alternative, contradictory perspective to educational policy on the placement of students in special education settings, which is a hotly debated topic in many countries throughout the world.

References

Abu-Hilal, M. M. (2001). Correlates of achievement in the United Arab Emirates: A sociocultural study. In D. M. McInerney & S. Van Etten (Eds.), *Research on sociocultural influences on motivation and learning* (Vol. 1, pp. 205–230). Greenwich, CT: Information Age.

Abu-Hilal, M. M., & Aal-Hussain, A. A. (1997). Dimensionality and hierarchy of the SDQ in a non-Western milieu: A test of self-concept invariance across gender. *Journal of Cross-Cultural Psychology*, 28, 535–553. doi:10.1177/0022022197285002

Abu-Hilal, M. M., & Bahri, T. M. (2000). Self-concept: The generalizability of research on the SDQ, Marsh/Shavelson model and I/E reference model to United Arab Emirates students. *Social Behavior and Personality*, 28, 309–322. doi:10.2224/sbp.2000.28.4.309

Alicke, M. D., Zell, E., & Bloom, D. L. (2010). Mere categorization and the frog-pond effect. *Psychological Science*, 21, 174–177. doi:10.1177/0956797609357718

Alwin, D. F., & Otto, L. B. (1977). High school context effects on aspirations. *Sociology of Education*, 50, 259–273. doi:10.2307/2112499

American Institutes for Research. (2005). *Reassessing U.S. international mathematics performance: New findings from the 2003 TIMSS and PISA*. Washington, DC: Author. Retrieved from http://www.air.org/files/TIMSS_PISA_math_study1.pdf

Bandura, A. (2006). Toward a psychology of human agency. *Perspectives on Psychological Science*, 1, 164–180. doi:10.1111/j.1745-6916.2006.00011.x

Bates, E. (1990). Language about me and you: Pronominal reference and the emerging concept of self. In D. Cicchetti & M. Beeghly (Eds.), *The self in transition: Infancy to childhood* (pp. 165–182). Chicago, IL: University of Chicago Press.

Beauducel, A., & Herzberg, P. Y. (2006). On the performance of maximum likelihood versus means and variance adjusted weighted least squares estimation in CFA. *Structural Equation Modeling*, 13, 186–203. doi:10.1207/s15328007sem1302_2

Bornholt, L. J. (1997). Aspects of self knowledge about activities with young children. *Every Child*, 3, 15–18.

Bouffard, T., Markovits, H., Vezeau, C., Boisvert, M., & Dumas, C. (1998). The relation between accuracy of self-perception and cognitive development. *British Journal of Educational Psychology*, 68, 321–330. doi:10.1111/j.2044-8279.1998.tb01294.x

Bruner, J. (1996). A narrative model of self construction. *Psyke & Logos*, 17, 154–170.

Chapman, J. W., & Tunmer, W. E. (1995). Development of children's reading self-concepts: An examination of emerging subcomponents and their relation with reading achievement. *Journal of Educational Psychology*, 87, 154–167. doi:10.1037/0022-0663.87.1.154

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling*, 14, 464–504. doi:10.1080/10705510701301834

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling*, 9, 233–255. doi:10.1207/S15328007SEM0902_5

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Damon, W., & Hart, D. (1988). *Self-understanding in childhood and adolescence*. New York, NY: Cambridge University Press.

Davis, J. (1966). The campus as a frog pond: An application of the theory of relative deprivation to career decisions of college men. *American Journal of Sociology*, 72, 17–31. doi:10.1086/224257

Davis-Kean, P. E., & Sandler, H. M. (2001). A meta-analysis of measures of self-esteem for young children: A framework for future measures. *Child Development*, 72, 887–906. doi:10.1111/1467-8624.00322

Diener, E. (2000). Subjective well-being: The science of happiness and a proposal for a national index. *American Psychologist*, 55, 34–43. doi:10.1037/0003-066X.55.1.34

Diener, E., & Fujita, F. (1997). Social comparison and subjective well-being. In B. P. Buunk & F. X. Gibbons (Eds.), *Health, coping, and well-being: Perspectives from social comparison theory* (pp. 329–358). Mahwah, NJ: Erlbaum.

DiStefano, C. (2002). The impact of categorization with confirmatory factor analysis. *Structural Equation Modeling*, 9, 327–346. doi:10.1207/S15328007SEM0903_2

Dolan, C. V. (1994). Factor analysis of variables with 2, 3, 5 and 7 response categories: A comparison of categorical variable estimators using simulated data. *British Journal of Mathematical and Statistical Psychology*, 47, 309–326. doi:10.1111/j.2044-8317.1994.tb01039.x

Dweck, C. S. (1999). *Self-theories: Their role in motivation, personality, and development*. Philadelphia, PA: Psychology Press.

Eccles, J. S. (with Adler, T. F., Futterman, R., Goff, S. B., Kaczala, C. M., Meece, J. L., & Midgley, C.). (1983). Expectancies, values, and academic behaviors. In J. T. Spence (Ed.), *Achievement and achievement motivation: Psychological and sociological approaches* (pp. 75–146). San Francisco, CA: Freeman.

Eccles, J., Wigfield, A., Harold, R. D., & Blumenfeld, P. (1993). Age and gender differences in children's self- and task perceptions during elementary school. *Child Development*, 64, 830–847. doi:10.2307/1131221

Eder, R. A., & Mangelsdorf, S. C. (1997). The emotional basis of early personality development: Implications for the emergent self-concept. In R. Hogan, J. Johnson, & S. Briggs (Eds.), *Handbook of personality psychology* (pp. 209–240). San Diego, CA: Academic Press. doi:10.1016/B978-012134645-4/50010-X

Enders, C. K., & Tofghi, D. (2007). Centering predictor variables in cross-sectional multilevel models: A new look at an old issue. *Psychological Methods*, 12, 121–138. doi:10.1037/1082-989X.12.2.121

Ertl, H. (2006). Educational standards and the changing discourse on education: The reception and consequences of the PISA study in Germany. *Oxford Review of Education*, 32, 619–634. doi:10.1080/03054980600976320

Fantuzzo, J. W., McDermott, P. A., Manz, P. H., Hampton, V. R., & Burdick, N. A. (1996). The pictorial scale of perceived competence and social acceptance: Does it work with low-income urban children? *Child Development*, 67, 1071–1084. doi:10.2307/1131880

Festinger, L. (1954). A theory of social comparison processes. *Human Relations*, 7, 117–140. doi:10.1177/001872675400700202

Fischer, K. W. (1980). A theory of cognitive development: The control and construction of hierarchies of skills. *Psychological Review*, 87, 477–531. doi:10.1037/0033-295X.87.6.477

Guay, F., Marsh, H. W., & Boivin, M. (2003). Academic self-concept and academic achievement: Developmental perspectives on their causal ordering. *Journal of Educational Psychology, 95*, 124–136. doi:10.1037/0022-0663.95.1.124

Harter, S. (1983). Developmental perspectives on the self-system. In P. H. Mussen (Ed.), *Handbook of child psychology* (Vol. 4, 4th ed., pp. 275–385). New York, NY: Wiley.

Harter, S. (1998). The development of self-representations. In W. Damon (Ed.) & S. Eisenberg (Vol. Ed.), *Handbook of child psychology: Vol. 3. Social, emotional, and personality development* (5th ed., pp. 553–617). New York, NY: Wiley.

Harter, S. (1999). *The construction of the self: A developmental perspective*. New York, NY: Guilford Press.

Harter, S. (2006). The self. In N. Eisenberg, W. Damon, & R. M. Lerner (Eds.), *Handbook of child psychology: Vol. 3. Social, emotional, and personality development* (6th ed., pp. 505–570). Hoboken, NJ: Wiley.

Harter, S. (2012). *The construction of the self: Developmental and socio-cultural foundations* (2nd ed.). New York, NY: Guilford Press.

Hattie, J. (1992). *Self-concept*. Hillsdale, NJ: Erlbaum.

Hattie, J., & Marsh, H. W. (1996). Future directions in self-concept research. In B. A. Bracken (Ed.), *Handbook of self-concept* (pp. 421–462). New York, NY: Wiley.

Helson, H. (1964). *Adaptation-level theory*. New York, NY: Harper & Row.

Hopmann, S., Brinek, G., & Retzl, M. (Eds.). (2007). *PISA According to PISA*. Vienna, Austria: Verlag.

Hu, L.-T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*, 1–55. doi:10.1080/10705519909540118

Huguet, P., Dumas, F., Marsh, H. W., Régner, I., Wheeler, L., Suls, J., . . . Nezlek, J. (2009). Clarifying the role of social comparison in the big-fish-little-pond effect (BFLPE): An integrative study. *Journal of Personality and Social Psychology, 97*, 156–170. doi:10.1037/a0015558

Hutchison, G., & Schagen, I. (2007). Comparisons between PISA and TIMSS—Are we the man with two watches? In T. Loveless (Ed.), *Lessons learned: What international assessments tell us about math achievement* (pp. 227–261). Washington, DC: Brookings Institution.

Hyman, H. (1942). The psychology of subjective status. *Psychological Bulletin, 39*, 473–474.

James, W. (1963). *The principles of psychology*. New York, NY: Holt, Rinehart & Winston. (Original work published 1890).

Jerusalem, M. (1984). Reference group, learning environment and self-evaluations: A dynamic multi-level analysis with latent variables. In R. Schwarzer (Ed.), *Advances in psychology: Vol. 21. The self in anxiety, stress and depression* (pp. 61–73). Amsterdam, the Netherlands: North-Holland. doi:10.1016/S0166-4115(08)62115-9

Kagen, S. L., Moore, E., & Bredekamp, S. (Eds.). (1995). *Reconsidering children's early development and learning: Toward common views and vocabulary* (Report No. 95-03). Washington, DC: National Education Goals Panel.

Kreft, I. G. G., de Leeuw, J., & Aiken, L. S. (1995). The effect of different forms of centering in hierarchical linear models. *Multivariate Behavioral Research, 30*, 1–21. doi:10.1207/s15327906mbr3001_1

Lewis, M., & Brooks-Gunn, J. (1979). *Social cognition and the acquisition of self*. New York, NY: Plenum Press. doi:10.1007/978-1-4684-3566-5

Liem, G. A. D., Marsh, H. W., Martin, A. J., McInerney, D. M., & Yeung, A. A. (2013). The big-fish-little-pond effect and a national policy of within-school ability streaming: Alternative frames of reference. *American Educational Research Journal, 50*, 326–370. doi:10.3102/0002831212464511

Lüdtke, O., Marsh, H. W., Robitzsch, A., & Trautwein, U. (2011). A 2 × 2 taxonomy of multilevel latent contextual models: Accuracy-bias trade-offs in full and partial error correction models. *Psychological Methods, 16*, 444–467. doi:10.1037/a0024376

Lüdtke, O., Marsh, H. W., Robitzsch, A., Trautwein, U., Asparouhov, T., & Muthén, B. (2008). The multilevel latent covariate model: A new, more reliable approach to group-level effects in contextual studies. *Psychological Methods, 13*, 203–229. doi:10.1037/a0012869

Marsh, H. W. (1974). *Judgmental anchoring: Stimulus and response variables* (Unpublished doctoral dissertation). University of California, Los Angeles.

Marsh, H. W. (1984). Self-concept: The application of a frame of reference model to explain paradoxical results. *Australian Journal of Education, 28*, 165–181.

Marsh, H. W. (1987). The big-fish-little-pond effect on academic self-concept. *Journal of Educational Psychology, 79*, 280–295. doi:10.1037/0022-0663.79.3.280

Marsh, H. W. (1989). Age and sex effects in multiple dimensions of self-concept: Preadolescence to early adulthood. *Journal of Educational Psychology, 81*, 417–430. doi:10.1037/0022-0663.81.3.417

Marsh, H. W. (1990). A multidimensional, hierarchical model of self-concept: Theoretical and empirical justification. *Educational Psychology Review, 2*, 77–172. doi:10.1007/BF01322177

Marsh, H. W. (1991). Failure of high ability schools to deliver academic benefits commensurate with their students' ability levels. *American Educational Research Journal, 28*, 445–480. doi:10.3102/00028312028002445

Marsh, H. W. (2007). *Self-concept theory, measurement and research into practice: The role of self-concept in educational psychology*. Leicester, England: British Psychological Society.

Marsh, H. W., Abduljabbar, A. S., Abu-Hilal, M. M., Morin, A. J. S., Abdelfattah, F., Leung, K. C., & Parker, P. (2013). Factorial, convergent, and discriminant validity of TIMSS math and science motivation measures: A comparison of Arab and Anglo-Saxon countries. *Journal of Educational Psychology, 105*, 108–128. doi:10.1037/a0029907

Marsh, H. W., Balla, J. R., & McDonald, R. P. (1988). Goodness-of-fit indices in confirmatory factor analysis: The effect of sample size. *Psychological Bulletin, 103*, 391–410. doi:10.1037/0033-2909.103.3.391

Marsh, H. W., Chesson, D., Craven, R. G., & Roche, L. (1995). The effects of gifted and talented programs on academic self-concept: The big fish strikes again. *American Educational Research Journal, 32*, 285–319. doi:10.3102/00028312032002285

Marsh, H. W., & Craven, R. (1997). Academic self-concept: Beyond the dustbowl. In G. Phye (Ed.), *Handbook of classroom assessment: Learning, achievement, and adjustment* (pp. 131–198). Orlando, FL: Academic Press.

Marsh, H. W., & Craven, R. G. (2006). Reciprocal effects of self-concept and performance from a multidimensional perspective: Beyond seductive pleasure and unidimensional perspectives. *Perspectives on Psychological Science, 1*, 133–163. doi:10.1111/j.1745-6916.2006.00010.x

Marsh, H. W., Craven, R. G., & Debus, R. (1998). Structure, stability, and development of young children's self-concepts: A multicohort-multioccasion study. *Child Development, 69*, 1030–1053. doi:10.1111/j.1467-8624.1998.tb06159.x

Marsh, H. W., Debus, R., & Bornholt, L. (2005). Validating young children's self-concept responses: Methodological ways and means to understand their responses. In D. M. Teti (Ed.), *Handbook of research methods in developmental science* (pp. 138–160). Oxford, England: Blackwell. doi:10.1002/9780470756676.ch8

Marsh, H. W., Ellis, L., & Craven, R. G. (2002). How do preschool children feel about themselves? Unraveling measurement and multidimensional self-concept structure. *Developmental Psychology, 38*, 376–393. doi:10.1037/0012-1649.38.3.376

Marsh, H. W., & Hau, K.-T. (2003). Big-fish-little-pond effect on academic self-concept: A cross-cultural (26-country) test of the negative

effects of academically selective schools. *American Psychologist*, 58, 364–376. doi:10.1037/0003-066X.58.5.364

Marsh, H. W., & Hau, K.-T. (2004). Explaining paradoxical relations between academic self-concepts and achievements: Cross-cultural generalizability of the internal–external frame of reference predictions across 26 countries. *Journal of Educational Psychology*, 96, 56–67. doi:10.1037/0022-0663.96.1.56

Marsh, H. W., Hau, K.-T., & Craven, R. G. (2004). The big-fish-little-pond effect stands up to scrutiny. *American Psychologist*, 59, 269–271. doi:10.1037/0003-066X.59.4.269

Marsh, H. W., Hau, K.-T., & Grayson, D. (2005). Goodness of fit in structural equation models. In A. Maydeu-Olivares & J. J. McArdle (Eds.), *Contemporary psychometrics: A festschrift for Roderick P. McDonald* (pp. 276–340). Mahwah, NJ: Erlbaum.

Marsh, H. W., Hau, K.-T., & Wen, Z. (2004). In search of golden rules: Comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing Hu and Bentler's (1999) findings. *Structural Equation Modeling*, 11, 320–341. doi:10.1207/s15328007sem1103_2

Marsh, H. W., Kong, C.-K., & Hau, K.-T. (2000). Longitudinal multilevel models of the big-fish-little-pond effect on academic self-concept: Counterbalancing contrast and reflected-glory effects in Hong Kong schools. *Journal of Personality and Social Psychology*, 78, 337–349. doi:10.1037/0022-3514.78.2.337

Marsh, H. W., Lüdtke, O., Robitzsch, A., Trautwein, U., Asparouhov, T., Muthén, B., & Nagengast, B. (2009). Doubly-latent models of school contextual effects: Integrating multilevel and structural equation approaches to control measurement and sampling error. *Multivariate Behavioral Research*, 44, 764–802. doi:10.1080/00273170903333665

Marsh, H. W., & O'Mara, A. (2010). Long-term total negative effects of school-average ability on diverse educational outcomes: Direct and indirect effects of the big-fish-little-pond effect. *Zeitschrift für Pädagogische Psychologie*, 24, 51–72. doi:10.1024/1010-0652.a000004

Marsh, H. W., & Parker, J. W. (1984). Determinants of student self-concept: Is it better to be a relatively large fish in a small pond even if you don't learn to swim as well? *Journal of Personality and Social Psychology*, 47, 213–231. doi:10.1037/0022-3514.47.1.213

Marsh, H. W., Seaton, M., Trautwein, U., Lüdtke, O., Hau, K. T., O'Mara, A. J., & Craven, R. G. (2008). The big-fish-little-pond-effect stands up to critical scrutiny: Implications for theory, methodology, and future research. *Educational Psychology Review*, 20, 319–350. doi:10.1007/s10648-008-9075-6

Marsh, H. W., Trautwein, U., Lüdtke, O., Baumert, J., & Köller, O. (2007). Big-fish-little-pond effect: Persistent negative effects of selective high schools on self-concept after graduation. *American Educational Research Journal*, 44, 631–669. doi:10.3102/0002831207306728

Marsh, H. W., & Yeung, A. S. (1997). Causal effects of academic self-concept on academic achievement: Structural equation models of longitudinal data. *Journal of Educational Psychology*, 89, 41–54. doi:10.1037/0022-0663.89.1.41

Marsh, H. W., & Yeung, A. S. (1999). The lability of psychological ratings: The chameleon effect in global self-esteem. *Personality and Social Psychology Bulletin*, 25, 49–64. doi:10.1177/0146167299025001005

Millsap, R. E. (2011). *Statistical approaches to measurement invariance*. New York, NY: Routledge.

Möller, J., Streblow, L., & Pohlmann, B. (2009). Achievement and self-concept of students with learning disabilities. *Social Psychology of Education*, 12, 113–122. doi:10.1007/s11218-008-9065-z

Morse, S., & Gergen, K. J. (1970). Social comparison, self-consistency, and the concept of self. *Journal of Personality and Social Psychology*, 16, 148–156. doi:10.1037/h0029862

Muthén, B. O., & Kaplan, D. (1985). A comparison of some methodologies for the factor analysis of non-normal Likert variables. *British Journal of Mathematical and Statistical Psychology*, 38, 171–189. doi:10.1111/j.2044-8317.1985.tb00832.x

Muthén, L. K., & Muthén, B. O. (2013). *Mplus user's guide*. Los Angeles, CA: Muthén & Muthén.

Nagengast, B., & Marsh, H. W. (2012). Big fish in little ponds aspire more: Mediation and cross-cultural generalizability of school-average ability effects on self-concept and career aspirations in science. *Journal of Educational Psychology*, 104, 1033–1053. doi:10.1037/a0027697

National Center for Education Statistics. (2008). *Comparing NAEP, TIMSS, and PISA in mathematics and science*. Retrieved from http://nces.ed.gov/timss/pdf/naep_timss_pisa_comp.pdf

Neidorf, T. S., Binkley, M., Gattis, K., & Nohara, D. (2006). *Comparing mathematics content in the National Assessment of Educational Progress (NAEP), Trends in International Mathematics and Science Study (TIMSS), and Program for International Student Assessment (PISA) 2003 assessments: Technical report* (NCES 2006-029). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics.

Nicholls, J. G. (1979). Development of perceptions of own attainment and causal attributions of success and failure in reading. *Journal of Educational Psychology*, 71, 94–99. doi:10.1037/0022-0663.71.1.94

Olson, J. F., Martin, M. O., & Mullis, I. V. S. (Eds.). (2008). *TIMSS 2007 Technical Report*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Parducci, A. (1995). *Happiness, pleasure, and judgment: The contextual theory and its applications*. Mahwah, NJ: Erlbaum.

Parker, P. D., Schoon, I., Tsai, Y.-M., Nagy, G., Trautwein, U., & Eccles, J. S. (2012). Achievement, agency, gender, and socioeconomic background as predictors of postschool choices: A multicontext study. *Developmental Psychology*, 48, 1629–1642. doi:10.1037/a0029167

Penn, C. S., Burnett, P. C., & Patton, W. (2001). The impact of attributional feedback on the self-concept of children aged four to six years in preschool. *Australian Journal of Guidance and Counselling*, 9, 21–34.

Raykov, T., & Marcoulides, G. A. (2004). Using the delta method for approximate interval estimation of parameter functions in SEM. *Structural Equation Modeling*, 11, 621–637. doi:10.1207/s15328007sem1104_7

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York, NY: Wiley. doi:10.1002/9780470316696

Ruble, D. N., & Dweck, C. S. (1995). Self-conceptions, person conceptions, and their development. In N. Eisenberg (Ed.), *Review of personality and social psychology: Vol. 15. Social development* (pp. 109–139). Thousand Oaks, CA: Sage.

Russell, L., Bornholt, L., & Ouvrier, R. (2002). Brief cognitive screening and self concepts for children with low intellectual functioning. *British Journal of Clinical Psychology*, 41, 93–104. doi:10.1348/01446502163831

Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. New York, NY: Chapman and Hall/CRC. doi:10.1201/9781439821862

Schwartz, S. H., & Bilsky, W. (1990). Toward a theory of the universal content and structure of values: Extensions and cross-cultural replications. *Journal of Personality and Social Psychology*, 58, 878–891. doi:10.1037/0022-3514.58.5.878

Seaton, M., & Marsh, H. W. (2013). Celebrating methodological–substantive synergy: Self-concept theory and methodological innovation. In D. McInerney, H. W. Marsh, R. G. Craven, & F. Guay (Eds.), *International advances in self research: Vol. 4. Theory driving research: New wave perspectives on self-processes and human development* (pp. 161–181). Greenwich, CT: Information Age Press.

Seaton, M., Marsh, H. W., & Craven, R. G. (2009). Earning its place as a pan-human theory: Universality of the big-fish-little-pond effect across 41 culturally and economically diverse countries. *Journal of Educational Psychology*, 101, 403–419. doi:10.1037/a0013838

Seaton, M., Marsh, H. W., & Craven, R. G. (2010). Big-fish-little-pond effect: Generalizability and moderation—Two sides of the same coin. *American Educational Research Journal*, 47, 390–433. doi:10.3102/0002831209350493

Seaton, M., Marsh, H. W., Dumas, F., Huguet, P., Monteil, J.-M., Régner, I., . . . Wheeler, L. (2008). In search of the big fish: Investigating the coexistence of the big-fish-little-pond effect with the positive effects of upward comparisons. *British Journal of Social Psychology*, 47, 73–103. doi:10.1348/014466607X202309

Segall, M. H., Lonner, W. J., & Berry, J. W. (1998). Cross-cultural psychology as a scholarly discipline: On the flowering of culture in behavioural research. *American Psychologist*, 53, 1101–1110. doi:10.1037/0003-066X.53.10.1101

Seligman, M. E. P., & Csikszentmihalyi, M. (2000). Positive psychology: An introduction. *American Psychologist*, 55, 5–14. doi:10.1037/0003-066X.55.1.5

Skaalvik, E. M., & Hagtvæt, K. A. (1990). Academic achievement and self-concept: An analysis of causal predominance in a developmental perspective. *Journal of Personality and Social Psychology*, 58, 292–307. doi:10.1037/0022-3514.58.2.292

Stipek, D., & Mac Iver, D. (1989). Developmental change in children's assessment of intellectual competence. *Child Development*, 60, 521–538. doi:10.2307/1130719

Stouffer, S. A., Suchman, E. A., DeVinney, L. C., Star, S. A., & Williams, R. M. (1949). *The American soldier: Adjustments during army life* (Vol. 1). Princeton, NJ: Princeton University Press.

Tymms, P. (2001). A test of the big fish in a little pond hypothesis: An investigation into the feelings of seven-year-old pupils in school. *School Effectiveness and School Improvement*, 12, 161–181. doi:10.1076/seesi.12.2.161.3452

Upshaw, H. S. (1969). The Personal Reference Scale: An approach to social judgment. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 4, pp. 315–370). New York, NY: Academic Press. doi:10.1016/S0065-2601(08)60081-7

Wedell, D. H., & Parducci, A. (2000). Social comparison: Lessons from basic research on judgment. In J. Suls & L. Wheeler (Eds.), *Handbook of social comparison: Theory and research* (pp. 223–252). Dordrecht, the Netherlands: Kluwer Academic. doi:10.1007/978-1-4615-4237-7_12

Wigfield, A., & Eccles, J. S. (1992). The development of achievement task values: A theoretical analysis. *Developmental Review*, 12, 265–310. doi:10.1016/0273-2297(92)90011-P

Wigfield, A., Eccles, J. S., Yoon, K. S., Harold, R. D., Arbreton, A. J. A., Freedman-Doan, C., & Blumenfeld, P. C. (1997). Change in children's competence beliefs and subjective task values across the elementary school years: A 3-year study. *Journal of Educational Psychology*, 89, 451–469. doi:10.1037/0022-0663.89.3.451

Wu, M. (2009). A comparison of PISA and TIMSS 2003 achievement results in mathematics. *Prospects*, 39, 33–46. doi:10.1007/s11125-009-9109-y

Zell, E., & Alicke, M. D. (2009). Contextual neglect, self-evaluation, and the frog-pond effect. *Journal of Personality and Social Psychology*, 97, 467–482. doi:10.1037/a0015453

Received July 13, 2013

Revision received March 3, 2014

Accepted May 8, 2014 ■

Social Consequences of Academic Teaming in Middle School: The Influence of Shared Course Taking on Peer Victimization

Leslie Echols
University of California, Los Angeles

This study examined the influence of academic teaming (i.e., sharing academic classes with the same classmates) on the relationship between social preference and peer victimization among 6th-grade students in middle school. Approximately 1,000 participants were drawn from 5 middle schools that varied in their practice of academic teaming. A novel methodology for measuring academic teaming at the individual level was employed, in which students received their own teaming score based on the unique set of classmates with whom they shared academic courses in their class schedule. On the basis of both peer- and self-reports of victimization, the results of 2 path models indicated that students with low social preference in highly teamed classroom environments were more victimized than low-preference students who experienced less teaming throughout the school day. This effect was exaggerated in higher performing classrooms. Implications for the practice of academic teaming were discussed.

Keywords: academic teaming, ability grouping, middle school, social preference, peer victimization

The early middle school years are rife with social and academic challenges as children make the move from elementary to secondary education. Many children experience decreases in school liking and engagement as they navigate the new middle school environment (Burchinal, Roberts, Zeisel, & Rowley, 2008)—an environment in which peer aggression is at its peak (Eslea et al., 2003; Seals & Young, 2003). Victims of such aggression are at heightened risk of academic difficulties (Erath, Flanagan, & Bierman, 2008; Juvonen, Wang, & Espinoza, 2011; Nakamoto & Schwartz, 2010; Schwartz, Gorman, Dodge, Pettit, & Bates, 2008); and may also suffer from a wide range of internalizing (depression, loneliness, low self-esteem) and externalizing (aggression, delinquency, poor self-regulation) symptoms (Hanish & Guerra, 2002; Hodges, Malone, & Perry, 1997; Schwartz et al., 2008).

In middle school, when fitting in and being accepted by peers are top social priorities (see Fournier, 2009), having low social preference among peers (i.e., being more disliked than liked) may increase the risk of peer victimization. Although low social preference may not always be associated with peer victimization, having low social preference among peers *and* being the victim of peer aggression consistently leads to the most negative adjustment outcomes (Hodges et al., 1997; Sandstrom & Cillessen, 2003). It is important, therefore, to understand the conditions under which low social preference among peers *does* contribute to peer victimization.

Previous research has documented certain individual characteristics (e.g., low impulse control) that make some low preference children vulnerable to peer victimization (cf. Sandstrom & Cillessen, 2003), but no studies to date have considered the role of

school context, or how instruction is organized, in the relation between low social preference and peer victimization. This is surprising given the recognition among scholars that school context may explain vulnerability to peer victimization when individual characteristics fail to do so (see Brown, 1996; Merten, 1996). For example, Kochenderfer-Ladd and Skinner (2002) posited that repeated exposure of aggressors to their targets due to placement in the same classroom and/or school may contribute to the stability of victimization across the school years. Brown (1996) likewise suggested that a restricted range of peer encounters at school (as opposed to mixing with a wider range of grade mates) may contribute to reputation formation and fewer opportunities for change. The purpose of this study, therefore, was to examine whether the way in which students are grouped together in middle school—defined as their likelihood of taking classes with the same classmates—contributes to victimization for children with low social preference among their peers.

Interdisciplinary Teaming in Middle School

In middle school, the extent to which children share their classes with the same classmates, and therefore the likelihood of repeated contact with aggressors, is often influenced by whether interdisciplinary teaming is practiced in their school. Interdisciplinary teaming consists of a core set of teachers responsible for teaching the same group of students—typically a subset of same-grade students in the school population—with the intended benefits of greater collaboration among teachers and greater community among teachers and students, particularly during the transition to middle school (Thompson & Homestead, 2004). These benefits are well documented in the literature. For example, past research demonstrates that students in interdisciplinary teams have higher scores on standardized achievement tests, are more academically engaged, and have greater feelings of school belonging (Boyer & Bishop, 2004; Flowers & Mertens, 2003; Flowers, Mertens, & Mulhall, 1999; Lee & Smith, 1993; Wallace, 2007).

This article was published Online First August 4, 2014.

Correspondence concerning this article should be addressed to Leslie Echols, Department of Education, UCLA, Box 951521, Los Angeles, CA 90095. E-mail: leslie.echols@ucla.edu

Although the practice of interdisciplinary teaming is widespread—estimated to be in use in nearly 80% of all U.S. middle schools (McEwin, Dickinson, & Jenkins, 2003)—little is known about the role of interdisciplinary teaming in children’s relationships with and treatment by their peers. Although students change classrooms and teachers each period when they are teamed, this practice may restrict their exposure to the general student body at their school because their classmates are always composed of members of their team. Social preference may therefore be determined largely by the reputations formed within their team. Popular or well-liked students may enjoy taking classes with many of the same classmates who regard them as high status members of the peer group, but peer-rejected or disliked students may suffer the consequences of being repeatedly exposed to classmates with whom they have negative social relationships. In other words, interdisciplinary teaming may be socially beneficial for high-preference children but detrimental for children with low social preference, who must endure a poor reputation throughout the majority of the school day.

Academic Teaming: A Special Case of Interdisciplinary Teaming

In middle schools in which interdisciplinary teaming is practiced, exposure to the same classmates throughout the school day may be influenced by the type and amount of interdisciplinary teaming that occurs. For example, in schools with a small number of teams relative to the size of the grade-level population, children may not have the same set of classmates each period, even though their classmates always come from the same pool (team) of students. On the other hand, in schools where there is a large number of teams relative to the size of the grade-level population (i.e., each team comprises only one classroom of students), interdisciplinary teaming would result in the same classmates traveling together from course to course for all of their academic classes—a special case of interdisciplinary teaming referred to here as *academic teaming*.

Unfortunately, the teaming literature does not differentiate between interdisciplinary teaming in general and the more specific case of academic teaming. In fact, one major limitation of previous research is that interdisciplinary teaming has been measured as a school-level dichotomous indicator (practiced/not practiced), making it virtually impossible to investigate individual outcomes associated with the extent of teaming that occurs. So although teaming may appear to have a positive effect on middle school adjustment for children overall, it is unclear whether there might be negative outcomes associated with the practice of teaming for some children, particularly those with low social preference among their peers.

Academic Teaming and Peer Victimization

Empirical research on social reputations indicates that peer status is less stable across changing peer settings than in settings in which peers remain the same (Bukowski & Newcomb, 1984; Coie & Kupersmidt, 1983). When considering the role of academic teaming in peer victimization, this research suggests that the relation between low social preference and victimization might be stronger when academic teaming is practiced and weaker when it

is not. To illustrate, for low-preference children in middle school, changing classes *and* classmates each period of the school day may help reduce their visibility among peers and their likelihood of being victimized because each class would be composed of a different set of classmates and social norms. In other words, children with low social preference who share the *fewest* number of classes with the same peers may have the *most* opportunities to avoid victimization. On the contrary, if the middle school structure is such that children take classes primarily with the same set of classmates, even when they change classrooms, social status hierarchies may be more salient to the peer group, increasing the probability that children with low social preference would also experience peer victimization.

Ability grouping. In many schools, interdisciplinary teams are composed of students with similar academic profiles, and students share all their classes with classmates performing at the same academic level (Ansalone, 2001, 2006; Dauber, Alexander, & Entwistle, 1996; Eccles, Midgley, & Wigfield, 1993; Oakes, 1981). Thus, in practice, teaming may be synonymous with ability grouping or academic tracking. In order to isolate the true effect of teaming, independent of ability grouping, it is therefore necessary to also consider the role of classroom academic performance (e.g., achievement level among classmates) in the relation between academic teaming, social preference, and peer victimization. The existing literature can help us understand how this unique set of individual and classroom characteristics may interact. For example, it is well documented that children who are performing well academically are more likely to be popular among (i.e., liked or accepted by) their peers (DeRosier, Kupersmidt, & Patterson, 1994; Guay, Boivin, & Hodges, 1999; Meijs, Cillessen, Scholte, Segers, & Spijker, 2010). As such, higher performing classrooms may be composed of a greater concentration of children with high social preference. It reasons that children with low social preference in such a context would be more likely to “stick out,” thus increasing their risk for victimization; this risk may be compounded if, due to academic teaming, these low-preference children remain with the same high-preference classmates throughout their academic schedule.

The Present Study

Although the academic benefits of interdisciplinary teaming in middle school are well understood, the social consequences associated with this common educational practice have been relatively unexplored. Certain types of interdisciplinary teaming, such as academic teaming, might increase the visibility of children’s reputations in the peer group. For high-preference children, this visibility could result in social benefits, but for low-preference children, academic teaming could make them more vulnerable to peer maltreatment, such as being victimized. The primary objective of this study, therefore, was to investigate the influence of academic teaming on the relation between social preference and the likelihood of victimization among peers. Because academic teaming is often practiced in conjunction with ability grouping, the next objective of this study was to examine whether classroom academic performance plays a role in the influence of academic teaming on this relation.

To achieve these objectives, some other limitations of the interdisciplinary-teaming literature were addressed. Rather than

rely on a school-level dichotomous indicator of teaming (practiced/not practiced), an individualized and continuous measure of teaming was developed for this study in order to account for the extent to which students share their classes with the same classmates across the academic subjects in their course schedule. Unlike much of the previous research that lacked a developmental analysis, this study was conducted with a large sample of sixth-grade students to capture the social effects of academic teaming during the transition year into middle school, when reputations and social hierarchies are being formed. To allow adequate time for these social processes to develop, the influence of academic teaming on the relation between social preference and peer victimization was examined in the spring of sixth grade (controlling for social preference and victimization in the fall). It was hypothesized that low social preference would be associated with greater victimization, especially for students who experienced greater academic teaming, thus being repeatedly exposed to the same classmates throughout the course of the school day.

Method

Participants

Participants were drawn from a larger sample of 5,076 sixth-graders across two cohorts of students participating in the UCLA Middle School Diversity Project (MSDP), a longitudinal study of middle school adjustment in ethnically diverse schools from Northern and Southern California. Students were enrolled in one of 20 schools that varied in ethnic composition. To reduce confounds of ethnic diversity with socioeconomic status (SES), schools at the extremes of the SES continuum were avoided; only schools within a 20–80% range of free and/or reduced-price meal (FRPM) eligibility were recruited for the study. At the time of this study, school records were available for 19 of the 20 schools.

As is typical for sixth-grade students in California middle schools, all students were enrolled in a different subject with a different teacher each period and rotated classrooms throughout the school day. However, the extent to which students' classmates differed or stayed the same from course to course varied by school. In many California middle schools, for example, the same group of classmates travels together from course to course, a practice referred to here as *academic teaming*. Using students' class schedules and the index of academic teaming described in detail below, participants were selected if they attended a school with significant within-school variability in academic teaming (i.e., some variation in classmates from course to course). Regardless of the extent of teaming practiced, however, students in these schools kept the same class schedule from fall to spring semester such that the classmates with whom they shared their courses remained the same throughout the academic year.

In order to examine whether high- or low teaming affects the relation between social preference and victimization, a subset of schools from the large sample was selected in which there was sufficient variance in the practice of teaming. Only five of the 19 schools for which class schedules were available met this criterion (i.e., the proportion of classmates that remained the same across all academic subjects ranged, on average, from .21 to .65). These five schools did not differ significantly from the overall sample in terms of FRPM eligibility or overall Academic Performance Index

(API) scores as reported by the California Department of Education (see Appendix). None of these schools housed special programs or magnet (e.g., gifted/highly gifted, science) centers. For two of the five schools that had substantial within-school variance in teaming scores, there was a significant correlation between academic teaming and classroom academic performance ($r = .67$ and $-.40$, respectively), suggesting that some schools may use teaming as a mechanism for ability grouping or academic tracking (e.g., grouping together remedial or honors students). The high ($M \geq .92$) average teaming scores for the remaining 14 schools in the larger sample demonstrate the prevalence of this middle school practice (see Appendix).

The analytic sample for the current study comprised 1,044 students (51.3% girls) from the 5 selected schools. The ethnic composition of the sample (based on student self-report) is as follows: 30.6% Latino/Mexican, 22.6% Asian (East/Southeast/South), 12.5% White, 11.9% African American, 3.0% Filipino/Pacific Islander, 14.2% multiethnic/biracial, and 5.2% other.

Procedure

Beginning in the fall of 2009, students with signed parental consent completed a questionnaire during a single period in one of their sixth-grade classes. Students recorded their answers independently as they followed instructions being read aloud by a graduate research assistant who reminded them of the confidentiality of their responses. A second researcher circulated around the classroom to help students as needed. This procedure was repeated (approximately 5 months later) in the spring semester of sixth grade. At both waves of data collection, students were given an honorarium of \$5 for completing the questionnaire.

Measures

Social preference. Social preference among peers was determined by peer nomination. In both the fall and spring of sixth grade, students were presented with a roster containing the names of all students in their grade level at their school, arranged by name (alphabetically by first name) and gender. Given the rotating structure of courses in California middle schools and the opportunity for interaction with many other grade mates throughout the school day, grade-level rosters were determined to be more appropriate than classroom-level rosters, which would have been limited to one set of peers to whom students were exposed in a given school day. Using the roster, students were instructed to record the names of their classmates in response to the questions, "Which sixth-grade students from your list would you like to hang out with at school?" and "Which sixth-grade students from your list do you *not* like to hang out with at school?" Students were allowed to record as many names as they desired but were instructed not to nominate themselves.

The conditional phrase "would you like to hang out with" was intentionally used as a measure of peer acceptance because it could include both whom students associated with already *and* whom they would like to hang out with if given the opportunity. Other peer acceptance measures commonly used in the literature (e.g., "who do you like the most at school?") also capture both established and desired associations with peers (cf. Lease & Axelrod, 2001; MacDonald & Cohen, 1995).

Similar to the procedure used by Coie, Dodge, and Coppotelli (1982), "not like" nominations received by each student were subtracted from "like" nominations received. Thus, social preference scores of 0 represented students who were equally liked and disliked, positive social preference scores (scores greater than 0) represented students who were liked more than they were disliked, and negative social preference scores (scores less than 0) represented students who were disliked more than they were liked.

Social impact. Social impact is often measured in conjunction with social preference in order to differentiate individuals with similar social preference scores who may be more or less known to members of peer group (Coie et al., 1982). For example, an individual who received five "like" nominations and five "not like" nominations may be more well known to peers than an individual who received one "like" nomination and one "not like" nomination, even though both individuals would be given a social preference score of 0. In other words, social impact is useful in detecting the strength of one's reputation (positive or negative). In order to control for the influence of reputation strength on peer victimization, social impact in the spring was calculated for each participant and used as a covariate in all analyses.

Victimization. Because social preference is a reputational measure of status among peers, peer reports of victimization may be more highly correlated with social preference than with self-reports of peer victimization. For this reason, both peer-reported and self-reported measures of victimization were used in this study.

Peer-reported victimization. On the same peer nomination measure as described above, students were instructed to record the names of their classmates in response to the question, "Which sixth-grade students from your list get picked on by other kids (get hit or pushed around, called bad names, talked about behind their backs)?" The total number of "picked on" nominations that each student received was then tallied to create a score of peer-reported victimization in both fall and spring of sixth grade.

Self-reported victimization. At each wave of data collection, students answered seven items about how often someone in their school had engaged in some type of aggression toward them (e.g., "hit, kicked, or pushed you," "called you bad names") since the beginning of the school year. Responses ranged from 1 (*never*) to 5 (*almost every day*). This new measure, created for the larger study, has been shown to relate to other indicators of social and emotional adjustment (see Lanza, Echols, & Graham, 2013). On the basis of high internal consistency in both the fall and spring of sixth grade ($\alpha = .86$ and $.87$, respectively), a mean of these items was computed and used as a single score of self-reported victimization.

Academic teaming. Students' class schedules were used to measure the proportion of participants' classmates who remained the same across all academic subjects during each semester. This proportion was calculated using an index of academic teaming that was created specifically for this study:

$$T = \frac{\sum \frac{C_x \cap C_y - 1}{C_x - 1}}{nP_2}, \text{ where } x = 1 \dots n, y = 1 \dots n, \text{ and } x \neq y.$$

Using the above formula, the proportion of classmates in each academic class (C_x) who were also in another academic class (C_y) was calculated for all possible academic course combinations in

each student's class schedule. The sum of these proportions was then divided by the total number of possible academic course combinations (nP_2) to create an average proportion of students in each participant's class schedule that remained the same throughout the academic subjects (i.e., math, science, English, social studies) in a given academic semester. The top half of the academic-teaming equation represents the overlap in classmates between two given academic courses (e.g., math and social studies) totaled across all possible course combinations (i.e., math and social studies, math and science, social studies and science, etc.). The bottom half of the equation represents the number of possible academic course combinations when each course is paired with every other course. Possible scores on this teaming index range between 0 and 1, with scores closer to 1 representing a higher proportion of students in one academic course who were also in every other academic course (i.e., complete academic teaming). For example, a score of .25 would indicate that 25% of a student's classmates remained the same across all four academic courses in his or her class schedule (low teaming), while a score of .75 would indicate that 75% of a student's classmates remained the same across all four academic courses (high teaming).

Classroom academic performance. Classroom academic performance was measured by average academic GPA among classmates according to the following procedure. First, on the basis of students' semester grades provided in school records, grade point average (GPA) was calculated for all participants for each academic course in their class schedule. Next, average GPA across classmates in each academic course was calculated. Finally, average classmate GPA in each course was averaged across the four academic courses in each participant's class schedule. Because the average academic performance to which students are exposed in their classrooms varies for students in middle school depending on their course schedules, each participant received an average classmate GPA score, ranging from 0 to 4, using the available school records data for participants.

Academic deviation. To control for risk of victimization associated with deviation from the norm for academic performance in students' academic courses, a difference score was calculated for each participant and used as a covariate in all analyses. To calculate this difference score, average classmate GPA was subtracted from average individual GPA for academic courses. Positive deviation scores represented students who were performing better than their classmates and negative deviation scores represented students who were performing worse than their classmates.

Planned Missing Design

In the larger study from which this sample was drawn, a three-form planned missing design was implemented in order to increase the efficiency of collecting data from such a large number of participants (see Graham, Taylor, Olchowski, & Cumsille, 2006). With this design, participants were given one of three questionnaires, each of which excluded a different set of measures, resulting in missing data on these measures for one third of participants. Because "missingness" was planned (i.e., under the control of the researchers) and not a function of other measured or unmeasured variables, these missing data were assumed to be missing completely at random (MCAR; see Little & Rubin, 1987). In the current study, only peer-reported victimization was part of the

planned missing design. There was a minimal amount (<4%) of unplanned missing data for this measure and all other measures in this study. Missing data were handled using full information maximum likelihood (FIML; described below).

Results

Descriptive statistics for all study variables are shown in Table 1. Correlations among study variables are shown in Table 2. Given the potential for reciprocal relations between social preference and victimization over time, a path (e.g., cross-lagged) model based on a structural equation modeling framework in Mplus (Muthén & Muthén, 2012) was used in order to measure the influence of social preference in the spring of sixth grade on victimization in the spring of sixth grade while simultaneously accounting for the influence of social preference and victimization in the fall of sixth grade. Two sets of models were estimated: one for peer-reported victimization and one for self-reported victimization. In each model, FIML was used in order to make use of all available data from participants. FIML is considered the most appropriate estimation technique for structural equation models when missing data are MCAR (Enders & Bandalos, 2001).

The results of the path models (described separately for peer- and self-reported victimization below) are shown in Table 3. Social preference and victimization in the fall and spring of sixth grade were modeled as observed variables. Individual path coefficients from covariates (gender, social impact, academic deviation) to observed variables are not shown, but all covariates were correlated with social preference and victimization in both fall and spring. In Step 1, academic teaming was included as a moderator of the relation between social preference and victimization in the spring of sixth grade, controlling for classroom academic performance. In Step 2, classroom academic performance was included in a three-way interaction term with academic teaming and social preference. As is standard practice when modeling higher order interaction terms, all lower order interaction terms (Social Preference \times Academic Teaming, Social Preference \times Classroom Academic Performance, Academic Teaming \times Classroom Academic Performance) were included in Step 2. R^2 values are reported for each model in order to evaluate the proportion of variance accounted for at each step. To ensure adequate sample size to detect close model fit, a separate power analysis for each model was

conducted following procedures outlined by MacCallum, Browne, and Sugawara (1996) for structural equation models (see Table 4). The power analyses confirmed sufficient sample size using even the most conservative criteria (power of .80 at $\alpha = .01$).

Peer-Reported Victimization

Social preference in the spring of sixth grade had a significant impact on peer-reported victimization in the spring of sixth grade, even after controlling for all relations between social preference and peer-reported victimization in the fall of sixth grade and between fall and spring of sixth grade. As shown in Step 1 of Table 3, the negative coefficient for this pathway indicates that as spring social preference increased, spring peer-reported victimization decreased; conversely, as spring social preference decreased, spring peer-reported victimization increased.

There was no main effect of academic teaming on peer-reported victimization in the spring. However, there was a significant interaction effect with social preference, indicating that the relation between social preference and peer-reported victimization in the spring was magnified by academic teaming. In other words, the more teaming students experienced, the greater the impact of social preference on peer-reported victimization. As shown in Figure 1, for students with high social preference, high teaming resulted in lower peer-reported victimization. For students with low social preference, however, high teaming resulted in particularly high peer-reported victimization. Thus, as hypothesized, academic teaming appears to be a risk factor for low-preference students at school.

As shown in Step 2 of Table 3, there was also a significant three-way interaction between social preference, academic teaming, and classroom academic performance; and the higher R^2 value indicates more variance in peer-reported victimization accounted for with this model. The three-way interaction is depicted in Figure 2. Each plotted slope shows the relation between social preference and peer-reported victimization at varying levels of academic teaming and classroom academic performance. Higher teaming resulted in greater victimization for children with low social preference in both higher and lower performing classrooms. In addition, children with low social preference had greater peer-reported victimization in higher compared with lower performing classes regardless of the amount of academic teaming they experienced.

Table 1
Descriptive Statistics for Study Variables

Variable	Min	Max	<i>M</i>	<i>SD</i>
Social preference fall 6th Grade	-19.00	20.00	2.50	3.35
Social preference spring 6th grade	-18.00	9.00	0.37	2.54
Social impact fall 6th grade	0.00	26.00	4.49	3.70
Social impact spring 6th grade	0.00	20.00	2.80	2.65
Peer-reported victimization fall 6th grade	0.00	9.00	0.40	0.89
Peer-reported victimization spring 6th grade	0.00	20.00	0.59	1.65
Self-reported victimization fall 6th grade	1.00	5.00	1.60	0.66
Self-reported victimization spring 6th grade	1.00	4.86	1.78	0.74
Academic teaming spring 6th grade	0.05	0.99	0.36	0.22
Classroom academic performance spring 6th grade	1.72	3.76	2.86	0.38
Academic deviation spring 6th grade	-2.79	2.19	0.08	0.79

Note. Min = minimum; Max = maximum; *M* = mean; *SD* = standard deviation.

Table 2
Correlation Among Study Variables

Variable	Fall social pref.	Spring social pref.	Fall peer-reported vic.	Spring peer-reported vic.	Fall self-reported vic.	Spring self-reported vic.	Academic teaming	Classroom academic performance	Academic deviation	Social impact
Fall social pref.	—	.546***	-.189***	-.235***	-.059	.016	-.038	.178***	.215***	.126***
Spring social pref.	—	-.302***	-.390***	-.098**	-.118**	-.053	.148***	.246***	-.172***	
Fall peer-reported vic.	—	.622***	.151***	.121**	.031	.009	-.012	.203***		
Spring peer-reported vic.	—	.170***	.208***	.032	.025	-.075*	.266***			
Fall self-reported vic.	—		.539***	.064*	-.097**	-.105**	.106**			
Spring self-reported vic.			—	.070	.002	-.104**	.173***			
Academic teaming					—		—	-.316***	.012	-.077*
Classroom academic performance						—	.079*	.093**		
Academic deviation							—			-.099**
Social impact									—	

Note. pref. = preference. vic. = victimhood. Academic teaming, classroom academic performance, academic deviation, and social impact were measured only in spring.

* $p < .05$. ** $p < .01$. *** $p < .001$.

Peer-reported victimization was greatest for children with low social preference in highly teamed, high performing classrooms.

Self-Reported Victimization

Similar to the model for peer-reported victimization, social preference in the spring of sixth grade had a significant impact (although smaller in magnitude) on self-reported victimization in the spring of sixth grade, even after controlling for all relations between social preference and self-reported victimization in the fall of sixth grade and between fall and spring of sixth grade. Again, the negative coefficient for this pathway indicates that as spring social preference increased, spring self-reported victimization decreased, and conversely that as spring social preference decreased, spring self-reported victimization increased.

In Step 1 of this model there was a significant main effect of academic teaming on self-reported victimization in the spring, indicating that students who experienced greater teaming reported more victimization. As shown in Figure 3, there was also a significant interaction effect with social preference such that for students with low social preference, greater teaming resulted in particularly high self-reported victimization. Similar to the model for peer-reported victimization, the relation between social preference and self-reported victimization was weakest when teaming was low. Unlike the model for peer-reported victimization, however, the three-way interaction between social preference, academic teaming, and classroom academic performance was not significant (see Step 2 of Table 3), indicating that the influence of academic teaming on the relation between social preference and

Table 3

Results of Path Models Testing the Moderating Role of Academic Teaming and Classroom Academic Performance on the Relation Between Social Preference and Victimization

Predictor	Peer-reported victimization		Self-reported victimization	
	Step 1 Est. (SE)	Step 2 Est. (SE)	Step 1 Est. (SE)	Step 2 Est. (SE)
Intercept (spring victimization)	.183 (.071)*	.127 (.043)**	.688 (.074)***	.662 (.077)***
Female	-.162 (.081)*	-.130 (.079)	-.022 (.050)	-.016 (.050)
Social impact	.082 (.016)***	.086 (.016)***	.033 (.010)**	.034 (.010)***
Academic deviation	.019 (.052)	.009 (.051)	-.042 (.032)	-.048 (.032)
Fall victimization	.987 (.047)***	.946 (.046)***	.606 (.037)***	.613 (.037)***
Fall social preference	-.027 (.015)	-.020 (.015)	.017 (.009)	.019 (.009)*
Spring social preference	-.136 (.020)***	-.142 (.020)***	-.025 (.012)*	-.026 (.013)*
Academic teaming	.275 (.182)	.627 (.199)**	.336 (.115)**	.324 (.131)*
Classroom academic performance	.259 (.109)*	.214 (.116)	.208 (.068)**	.243 (.073)**
Spring Social Preference \times Academic Teaming	-.240 (.084)**	-.385 (.085)***	-.145 (.051)**	-.178 (.057)**
Spring Social Preference \times Classroom Academic Performance		-.164 (.045)***		-.038 (.029)
Academic Teaming \times Classroom Academic Performance		1.386 (.478)**		-.219 (.302)
Spring Social Preference \times Academic Teaming \times Classroom Academic Performance		-.661 (.196)**		-.105 (.125)
<i>R</i> ²	.471	.488	.379	.379

Note. Est. = estimate. SE = standard error.

* $p < .05$. ** $p < .01$. *** $p < .001$.

Table 4

N for Test of Close Fit at Power = .80 For $\alpha = .01, .05$, and $.10$ With Varying Degrees of Freedom

Model	$\alpha = .01$	$\alpha = .05$	$\alpha = .10$
Step 1 (11 <i>df</i>)	834.38	612.50	504.69
Step 2 (32 <i>df</i>)	420.31	315.63	264.06

Note. *df* = degrees of freedom. Greater degrees of freedom reduce required sample size (see MacCallum, Browne, & Sugawara, 1996).

self-reported victimization was the same regardless of the level of academic performance in children's classes.

To summarize these results, for both peer- and self-reported victimization, as social preference increased, victimization decreased. Likewise, as social preference decreased, victimization increased. Predictably because of informant overlap, this effect appeared to be stronger for peer-reported than self-reported victimization. The interaction between social preference and academic teaming was significant for both types of victimization, and the relation between low social preference and victimization was greater when teaming was high. For peer-reported victimization, the relation between low social preference and victimization was greatest when both academic teaming and classroom academic performance were high.

Discussion

In early adolescence, perhaps more so than in any other time in development, status among peers contributes largely to children's social and emotional well-being and their overall adjustment in school (Wentzel, 2003). With many of these children using aggression to gain status (Pellegrini, 2002; Pellegrini & Long, 2002), having low status makes some children particularly vulnerable to

peer victimization (Sandstrom & Cillessen, 2003). Certain educational practices determine the type and amount of exposure children have to others in the peer group at school, which may affect their visibility as either high- or low-status members, further influencing their likelihood of being victimized. In particular, academic teaming influences the extent to which classmates remain the same from class to class throughout the school day, which could make social status more or less salient to their peers. In this study, social preference was used as the measure of status among peers, and the moderating role of academic teaming on the association between social preference and peer victimization for children in the sixth grade was examined.

Consistent with past research, the results indicated a significant negative relation between social preference and peer victimization. Even after accounting for reciprocal relations among these variables over time, lower social preference scores were associated with greater peer victimization in the spring of sixth grade according to both peer- and self-report. Academic teaming also had a significant negative effect on peer victimization, but only self-reported victimization, suggesting that the more children shared their classes with the same classmates, the more they perceived being victimized. For both peer- and self-reported victimization, academic teaming moderated the relation between social preference and victimization, increasing the risk of victimization among low preference peers. That is, regardless of the victimization measure that was used, low-preference children in highly teamed classes were more victimized than low-preference children who shared fewer classmates throughout the school day.

There was an opposite effect of academic teaming on peer-reported victimization for children with high social preference. Although high-preference children were at decreased risk of peer victimization overall, this was especially true for high-preference children who experienced greater academic teaming. These results

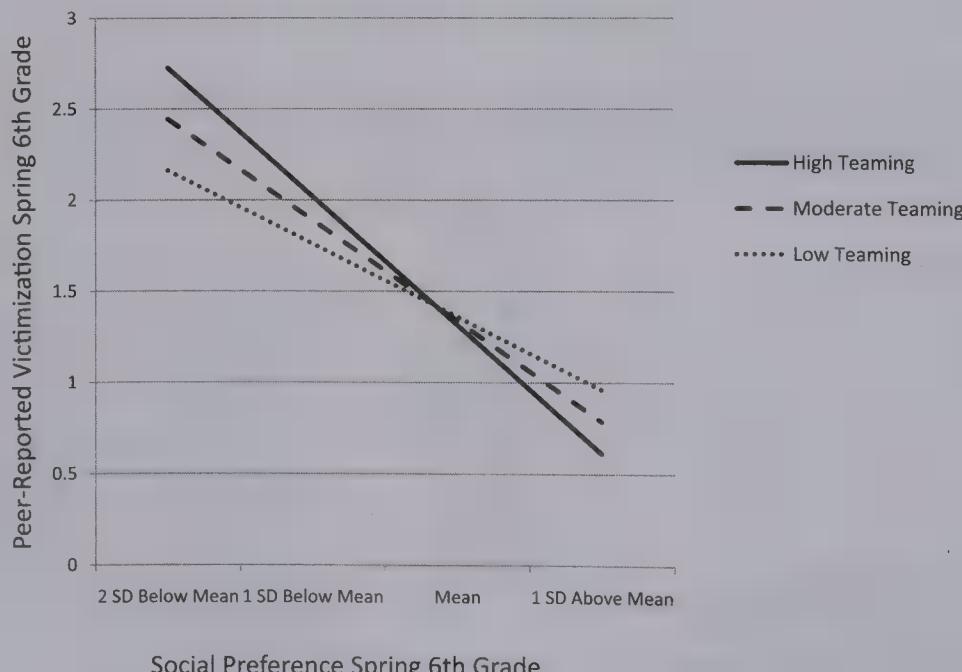


Figure 1. The moderating role of academic teaming on the relation between social preference and peer-reported victimization in middle school. Low teaming = 25% shared classmates, moderate teaming = 50% shared classmates, high teaming = 75% shared classmates. *SD* = standard deviation.

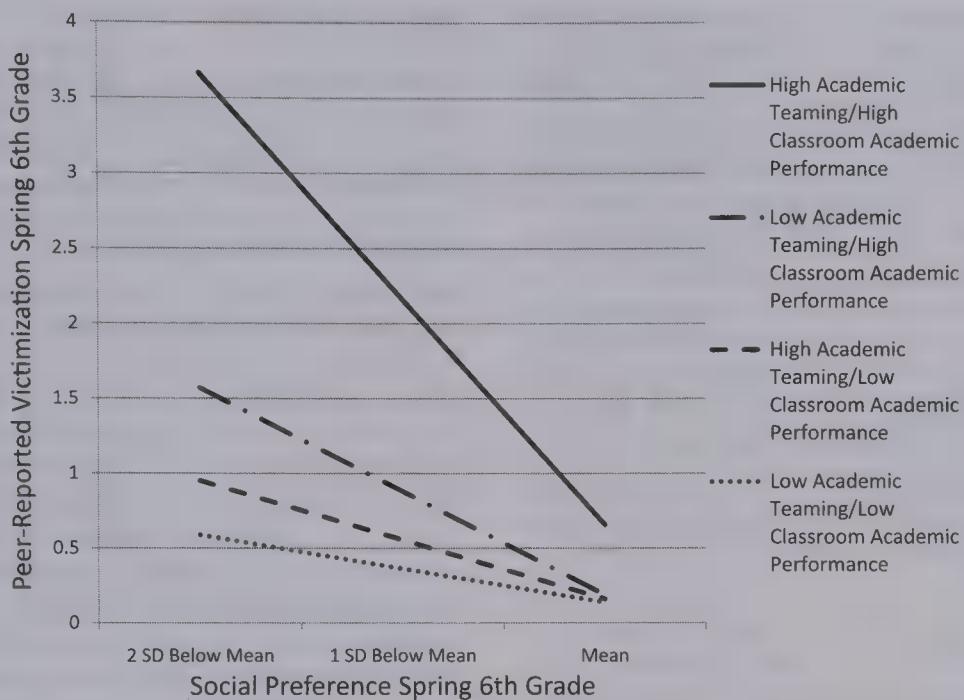


Figure 2. The moderating role of academic teaming and classroom academic performance on the relation between social preference and peer-reported victimization in middle school. Low teaming = 25% shared classmates, high teaming = 75% shared classmates. Low classroom academic performance = 1 SD below mean, high classroom academic performance = 1 SD above mean. *SD* = standard deviation.

support the hypothesis that academic teaming may increase the social visibility of children to their peers, which may be a promotive factor for children who enjoy high social preference in the peer group but a risk factor for low-preference children who rarely get the chance during the school day to escape their reputation.

The influence of academic teaming on the relation between social preference and peer victimization was further moderated by

classroom academic performance, such that children with low social preference were at greatest risk for victimization when they experienced higher levels of academic teaming *and* were taking classes with higher performing classmates. This effect was only observed for peer-reported victimization, suggesting that children with low social preference may not actually experience more victimization in higher performing classrooms but may stand out

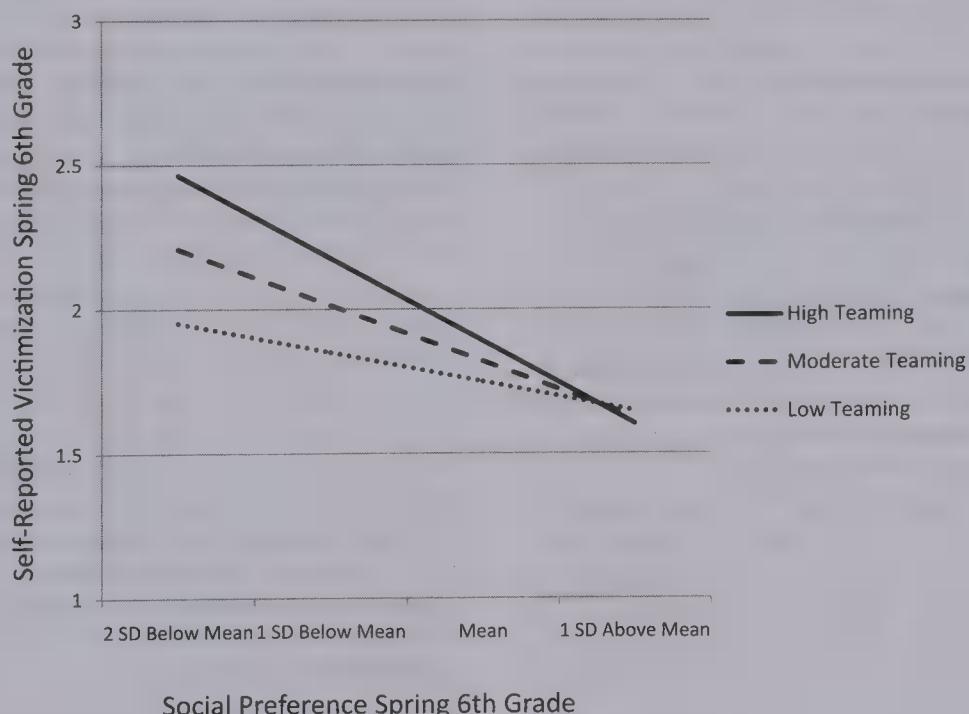


Figure 3. The moderating role of academic teaming on the relation between social preference and self-reported victimization in middle school. Low teaming = 25% shared classmates, moderate teaming = 50% shared classmates, high teaming = 75% shared classmates. *SD* = standard deviation.

more as victims compared with their higher preference peers. Because higher performing classrooms may be composed of more children with high social preference compared with average or lower performing classrooms (see Meijs et al., 2010; Newcomb, Bukowski, & Pattee, 1993), children who deviate from the norm of high social preference may be more visible to the peer group and therefore more easily identified as victims, especially if they are taking classes with the same classmates throughout the school day. Because social reputations are often difficult to change, especially in contexts in which peers remain the same (Brown, 1996), low-preference children in highly teamed, higher performing classrooms may be at risk for chronic victimization and a host of adjustment difficulties that could follow. Future research should consider the role of academic teaming on the unique social trajectories of children in higher performing classrooms who are at opposite ends of the social status hierarchy.

Strengths and Limitations

This study makes important contributions to the literature on both peer victimization and interdisciplinary teaming. By comparing the influence of social preference across both peer- and self-reported victimization, this study demonstrates that the relation between social preference and peer victimization may be evident regardless of how victimization is measured. In addition, in this study it was suggested that social visibility is the mechanism through which academic teaming influences the relation between low social preference and peer victimization, especially in higher performing classrooms. Although social visibility was not directly measured, it was assumed that repeated exposure to the same classmates through the practice of academic teaming would increase visibility among peers. This study introduces a novel methodological tool for measuring exposure in the peer group, but future research should consider other approaches to measuring social visibility (e.g., being a member of a particular "crowd").

Most notably, this is the first study in which interdisciplinary teaming has been distinguished from academic teaming in order to measure the extent to which children shared their classes with the same classmates throughout the school day. Because measuring academic teaming at the *individual* level and as a *continuous* variable is the only way to investigate individual outcomes associated with teaming, this study provides an important first look at the negative social consequences that may result from this common educational practice. Given the heightened risk of poor adjustment in middle school for children who are victimized, the prevalence of this school practice is alarming. However, it should be noted that children may share their classes with many of the same classmates even when interdisciplinary teaming as an instructional practice is *not* being utilized. Thus, the practice of academic teaming may or may not be synonymous with the practice of interdisciplinary teaming in all schools. As such, there may not be social risk associated with interdisciplinary teaming per se, but, rather, the risk may only reside in repeated exposure to classmates. Future research should further differentiate between the practices of interdisciplinary and academic teaming and consider other individual and social risk factors that may make children more or less likely to benefit from teaming in all its various forms.

Future Directions

This study sets the stage for other important research on peer victimization and the measurement of classroom context in middle and high school education. In the present study, the overlap in peer- and self-reported victimization was not accounted for (i.e., self-reported victimization was not included as a covariate in the model for peer-reported victimization and vice versa). However, it may be that children with high victimization scores on one measure were not necessarily the same children with high victimization scores on the other measure. If there are indeed different subgroups of victims in middle school, as suggested by Sandstrom and Cillessen (2003), it is possible that the same feature of the classroom context might affect members of these subgroups in different ways. For example, children who perceive victimization but are not identified as victims by their peers may suffer from a victim mentality that is neither explained nor influenced by their classroom context, whereas children who are identified as victims by their peers and themselves report being victimized may be particularly vulnerable when their classmates stay the same from course to course throughout the school day. Although this "comorbidity" effect was not examined in the present study, future research might consider whether using these multiple informants has implications for assessing the risks associated with peer- versus self-perceived victimization in certain classroom contexts.

In this study, students' class schedules were used to create individualized measures of classroom characteristics (e.g., classroom academic performance). This appears to be a promising new approach to measuring classroom context for students in middle- and secondary education settings that has some noteworthy advantages. First, this method makes it possible to detect differences in the influence of classroom context at various levels of measurement: between classrooms and schools, between students depending on their course schedules, and even between courses taken by the same student. Next, this method may remove the need for multilevel modeling if nearly all the variance in classroom context resides between students (within schools) and not between classrooms or schools. When multilevel modeling is necessary, classroom context measured at the individual level may substantially increase the number of Level 2 units (e.g., if classroom characteristics across courses are nested within individuals nested within schools). Most important, with this method, the individual experiences of children in middle and high school can be understood in ways never before examined. Instead of relying on measures of context specific to one classroom or school, this method makes it possible to investigate context across classrooms specific to each child. In other words, the entire school day as experienced by individual children as they travel from class to class can now be observed. Although the primary contribution of this study is the substantive understanding of the role of academic teaming in schoolchildren's social adjustment in school, it is the hope that this novel approach to measuring classroom context will also make a significant methodological contribution to the literature.

Implications for Practice

Although interdisciplinary teaming may lead to some positive outcomes such as greater feelings of belonging in school, academic teaming may come with certain social costs that outweigh these benefits. During a time when status among peers is critical to

overall adjustment and the risk of being victimized is so high, researchers and practitioners would do well to consider the extent to which this practice should be used, particularly for vulnerable children in the peer group. For example, a less-restrictive teaming structure (e.g., large-enough teams, so that students are not required to share all their academic courses with the same classmates) might provide the academic benefits of this practice while avoiding the social costs.

A relatively simple intervention strategy for reducing victimization among children with low social preference in the peer group would involve scheduling their courses in a way that would provide them with maximum exposure to a diverse set of peers. Because school counselors are often responsible for course scheduling, one important topic for future research is the role that school counselors play in addressing the academic and social needs of their students through course scheduling practices.

When a teaming structure is imposed by the school or district, it might also be important to consider how teachers in their individual classrooms might organize instruction in order to minimize the negative impact of this practice on students with low social preference among their peers. For example, are students further clustered together (e.g., in the case of small group instruction) with a particular set of classmates, or do they have the opportunity to interact with a variety of students in class? Is seating by student choice, or do teachers implement seating charts? As both these factors could influence the extent of exposure of low-preference students to the same classmates, these are important topics for future research.

Students' social and academic lives are interrelated and are closely tied to their overall adjustment in school; it is therefore important to consider both the academic and social ramifications of any instructional practice. Until now, only the academic benefits of teaming have been considered. However, because interdisciplinary teaming, in general, and academic teaming, in particular, have a direct impact on the type and extent of social contact that children experience, the practice of teaming may be especially relevant to children's social adjustment in school. It is the hope that the findings reported here will stimulate other research on the benefits and risks associated with the practice of teaming for children in middle school and that the methodology used here will make it possible to examine whether such outcomes apply to all, or just some, children in the classroom.

References

Ansalone, G. (2001). Schooling, tracking, and inequality. *Journal of Children & Poverty*, 7, 33–47. doi:10.1080/10796120120038028

Ansalone, G. (2006). Tracking: A return to Jim Crow. *Race, Gender, & Class*, 13, 144–153.

Boyer, A. J., & Bishop, P. A. (2004). Young adolescent voices: Students' perceptions of interdisciplinary teaming. *Research in Middle Level Education Online*, 28. Retrieved from <http://www.amle.org/ServicesEvents/ResearchinMiddleLevelEducationOnline/tabid/173/Default.aspx>

Brown, B. B. (1996). Visibility, vulnerability, development, and context: Ingredients for a fuller understanding of peer rejection in adolescence. *The Journal of Early Adolescence*, 16, 27–36. doi:10.1177/0272431696016001002

Bukowski, W. M., & Newcomb, A. F. (1984). The stability and determinants of sociometric status and friendship choice: A longitudinal perspective. *Developmental Psychology*, 20, 941–952. doi:10.1037/0012-1649.20.5.941

Burchinal, M. R., Roberts, J. E., Zeisel, S. A., & Rowley, S. J. (2008). Social risk and protective factors for African American children's academic achievement and adjustment during the transition to middle school. *Developmental Psychology*, 44, 286–292. doi:10.1037/0012-1649.44.1.286

Coie, J. D., Dodge, K. A., & Coppotelli, H. (1982). Dimensions and types of social status: A cross-age perspective. *Developmental Psychology*, 18, 557–570. doi:10.1037/0012-1649.18.4.557

Coie, J. D., & Kupersmidt, J. B. (1983). A behavioral analysis of emerging social status in boys' groups. *Child Development*, 54, 1400–1416. doi:10.2307/1129803

Dauber, S. L., Alexander, K. L., & Entwistle, D. R. (1996). Tracking and transitions through the middle grades: Channeling educational trajectories. *Sociology of Education*, 69, 290–307. doi:10.2307/2112716

DeRosier, M. E., Kupersmidt, J. B., & Patterson, C. J. (1994). Children's academic and behavioral adjustment as a function of the chronicity and proximity of peer rejection. *Child Development*, 65, 1799–1813. doi:10.2307/1131295

Eccles, J. S., Midgley, C., & Wigfield, A. (1993). Development during adolescence: The impact of stage—environment fit on young adolescents' experiences in schools and in families. *American Psychologist*, 48, 90–101. doi:10.1037/0003-066X.48.2.90

Enders, C. K., & Bandalos, D. L. (2001). The relative performance of full information maximum likelihood estimation for missing data in structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal*, 8, 430–457.

Erath, S. A., Flanagan, K. S., & Bierman, K. L. (2008). Early adolescent school adjustment: Associations with friendship and peer victimization. *Social Development*, 17, 853–870. doi:10.1111/j.1467-9507.2008.00458.x

Eslea, M., Menesini, E., Morita, Y., O'Moore, M., Mora-Merchán, J. A., Pereira, B., & Smith, P. (2004). Friendship and loneliness among bullies and victims: Data from seven countries. *Aggressive Behavior*, 30, 71–83. doi:10.1002/ab.20006

Flowers, N., & Mertens, S. B. (2003). Middle school practices improve student achievement in high poverty schools. *Middle School Journal*, 35, 33–43.

Flowers, N., Mertens, S. B., & Mulhall, P. F. (1999). The impact of teaming: Five research-based outcomes. *Middle School Journal*, 31, 57–60.

Fournier, M. A. (2009). Adolescent hierarchy formation and the social competition theory of depression. *Journal of Social and Clinical Psychology*, 28, 1144–1172. doi:10.1521/jscp.2009.28.9.1144

Graham, J. W., Taylor, B. J., Olchowski, A. E., & Cumsille, P. E. (2006). Planned missing data designs in psychological research. *Psychological Methods*, 11, 323–343.

Guay, F., Boivin, M., & Hodges, E. V. E. (1999). Predicting change in academic achievement: A model of peer experiences and self-system processes. *Journal of Educational Psychology*, 91, 105–115. doi:10.1037/0022-0663.91.1.105

Hanish, L. D., & Guerra, N. G. (2002). A longitudinal analysis of patterns of adjustment following peer victimization. *Development and Psychopathology*, 14, 69–89. doi:10.1017/S0954579402001049

Hodges, E. V. E., Malone, M. J., & Perry, D. G. (1997). Individual risk and social risk as interacting determinants of victimization in the peer group. *Developmental Psychology*, 33, 1032–1039. doi:10.1037/0012-1649.33.6.1032

Juvonen, J., Wang, Y., & Espinoza, G. (2011). Bullying experiences and compromised academic performance across middle school grades. *The Journal of Early Adolescence*, 31, 152–173. doi:10.1177/0272431610379415

Kochenderfer-Ladd, B., & Skinner, K. (2002). Children's coping strategies: Moderators of the effects of peer victimization? *Developmental Psychology*, 38, 267–278. doi:10.1037/0012-1649.38.2.267

Lanza, H. I., Echols, L., & Graham, S. (2013). Deviating from the norm: Body mass index (BMI) differences and psychosocial adjustment among early adolescent girls. *Journal of Pediatric Psychology*, 38, 376–386. doi:10.1093/jpepsy/jss130

Lease, A. M., & Axelrod, J. L. (2001). Position of the peer group's perceived organizational structure: Relation to social status and friendship. *The Journal of Early Adolescence*, 21, 377–404. doi:10.1177/0272431601021004001

Lee, V. E., & Smith, J. B. (1993). Effects of school restructuring on the achievement and engagement of middle-grades students. *Sociology of Education*, 66, 164–187. doi:10.2307/2112735

Little, R. J. A., & Rubin, D. B. (1987). *Statistical analysis with missing data*. New York, NY: Wiley.

MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods*, 1, 130–149. doi:10.1037/1082-989X.1.2.130

MacDonald, C. D., & Cohen, R. (1995). Children's awareness of which peers like them and which peers dislike them. *Social Development*, 4, 182–193. doi:10.1111/j.1467-9507.1995.tb00059.x

McEwin, C. K., Dickinson, T. S., & Jenkins, D. M. (2003). *America's middle schools in the new century: Status and progress*. Westerville, OH: National Middle School Association.

Meijs, N., Cillessen, A. H. N., Scholte, R. H. J., Segers, E., & Spijkerman, R. (2010). Social intelligence and academic achievement as predictors of adolescent popularity. *Journal of Youth and Adolescence*, 39, 62–72. doi:10.1007/s10964-008-9373-9

Merten, D. E. (1996). Visibility and vulnerability: Responses to rejection by nonaggressive junior high school boys. *The Journal of Early Adolescence*, 16, 5–26. doi:10.1177/0272431696016001001

Muthén, L. K., & Muthén, B. O. (2012). *Mplus user's guide* (7th ed.). Los Angeles, CA: Muthén & Muthén.

Nakamoto, J., & Schwartz, D. (2010). Is peer victimization associated with academic achievement? A meta-analytic review. *Social Development*, 19, 221–242. doi:10.1111/j.1467-9507.2009.00539.x

Newcomb, A. F., Bukowski, W. M., & Pattee, L. (1993). Children's peer relations: A meta-analytic review of popular, rejected, neglected, controversial, and average sociometric status. *Psychological Bulletin*, 113, 99–128. doi:10.1037/0033-2909.113.1.99

Oakes, J. (1981). *Tracking policies and practices: School by school summaries. A study of schooling* (Technical Report No. 25), Los Angeles, CA: University of California, Graduate School of Education.

Pellegrini, A. D. (2002). Affiliative and aggressive dimensions of dominance and possible functions during early adolescence. *Aggression and Violent Behavior*, 7, 21–31. doi:10.1016/S1359-1789(00)00033-1

Pellegrini, A. D., & Long, J. D. (2002). A longitudinal study of bullying, dominance, and victimization during the transition from primary school through secondary school. *British Journal of Developmental Psychology*, 20, 259–280. doi:10.1348/026151002166442

Sandstrom, M. J., & Cillessen, A. H. N. (2003). Sociometric status and children's peer experiences: Use of the daily diary method. *Merrill-Palmer Quarterly*, 49, 427–452. doi:10.1353/mpq.2003.0025

Schwartz, D., Gorman, A. H., Dodge, K. A., Pettit, G. S., & Bates, J. E. (2008). Friendships with peers who are low or high in aggression as moderators of the link between peer victimization and declines in academic functioning. *Journal of Abnormal Child Psychology*, 36, 719–730. doi:10.1007/s10802-007-9200-x

Seals, D., & Young, J. (2003). Bullying and victimization: Prevalence and relationship to gender, grade level, ethnicity, self-esteem, and depression. *Adolescence*, 38, 735–747.

Thompson, K. F., & Homestead, E. R. (2004). Middle school organization through the 1970s, 1980s, and 1990s. *Middle School Journal*, 35, 56–60.

Wallace, J. J. (2007). Effects of interdisciplinary teaching team configuration upon the social bonding of middle school students. *Research in Middle Level Education*, 30, 1–18.

Wentzel, K. R. (2003). Sociometric status and adjustment in middle school: A longitudinal study. *The Journal of Early Adolescence*, 23, 5–28. doi:10.1177/0272431602239128

Appendix
School Characteristics of MSDP Schools

School	FRPM	API	Academic Teaming
1	0.42	807	0.21
2	0.31	832	0.31
3	0.80	850	0.39
4	0.68	704	0.43
5	0.54	708	0.65
6	0.56	825	0.92
7	0.29	846	0.93
8	0.72	650	0.93
9	0.29	836	0.94
10	0.67	810	0.94
11	0.68	658	0.95
12	0.50	755	0.95
13	0.57	757	0.95
14	0.77	806	0.96
15	0.38	838	0.96
16	0.21	889	0.97
17	0.72	681	0.99
18	0.45	831	0.99
19	0.43	839	1.00

Note. MSDP = Middle School Diversity Project; FRPM = free and reduced-price meal eligibility; API = Academic Performance Index. Academic teaming scores were based on average academic teaming experienced by participants in the same school. Schools 1–5 were used in the analyses reported in this study.

Received June 7, 2013
 Revision received April 25, 2014
 Accepted June 15, 2014 ■

Long-Term Implications of Early Education and Care Programs for Australian Children

Rebekah Levine Coley, Caitlin McPherran Lombardi, and Jacqueline Sims
Boston College

Using nationally representative data from the Longitudinal Study of Australian Children (LSAC; $N = 5,107$), this study assessed prospective connections between children's early education and care (EEC) experiences from infancy through preschool and their cognitive and behavioral functioning in 1st grade. Incorporating 6 waves of data, analyses found that greater duration and intensity of exposure to center EEC settings predicted heightened fluid intelligence but also decreased behavioral functioning across multiple realms and reporters. Assessment of the timing of exposure found that the combination of infant/toddler and preschool center EEC, rather than only preschool EEC, drove these patterns. Results largely replicate patterns from U.S. studies, suggesting the importance of identifying EEC programs and models that can support children's behavioral as well as cognitive skills. In contrast to U.S. results, associations between center EEC and children's later functioning did not extend to basic academic skills and were not moderated by family socioeconomic resources or child temperament.

Keywords: child care, early childhood education, school readiness, international comparison, propensity score weighting

Early education and care (EEC) programs serve diverse needs for children and families, from promoting children's cognitive and behavioral skills, to supporting parental employment, to promoting equality and cultural norms. As such, governments across many countries are directing resources toward accessible, affordable, and high quality early education and care programs, using diverse policy levers such as quality regulations, federal or state subsidies, and family tax breaks (Organization for Economic Cooperation and Development [OECD], 2006). An extensive body of research has assessed how attending EEC programs affects children's core cognitive and behavioral functioning. This research has consistently found that children who attend early education programs, particularly center-based programs in the year or two prior to kindergarten, show enhanced growth in cognitive skills in com-

parison to their peers. At the same time, research also has suggested that center EEC, particularly when full-time and begun early in life, may be detrimental for later behavioral functioning (e.g., Coley, Votruba-Drzal, Miller, & Koury, 2013; Magnuson, Ruhm, & Waldfogel, 2007; NICHD Early Child Care Research Network [ECCRN], 2003; Phillips, McCartney, & Sussman, 2006). The vast majority of the literature on EEC derives from U.S. studies. Although findings have been quite robust across numerous longitudinal data sets of children from the United States, there has been little replication across other countries (although see, e.g., Côté, Borge, Geoffroy, Rutter, & Tremblay, 2008; Geoffroy et al., 2010). As such, we have limited knowledge concerning other policy models of EEC and whether effects of EEC on children's cognitive and behavioral skills are generalizable to different populations and in diverse policy and cultural environments.

With a relatively similar economic and cultural context yet differences in EEC policy and use, Australia offers an interesting context to replicate this research. In this article we assessed links between the duration, intensity, and timing of center-based EEC from infancy through preschool and children's cognitive and behavioral skills in first grade in a nationally representative sample of Australian children. We further addressed whether such associations differed across children from more or less advantaged home environments, considering family income, parent educational attainment, and enriching home environments, and across children with easier versus more difficult temperaments. This work presents one of the only assessments of the long-term implications of EEC in Australia. Further, it seeks to expand the broader literature on EEC and children's functioning by contrasting the roles of the duration, intensity, and timing of exposure to center-based EEC programs using rigorous statistical models to help adjust for selection bias and unmeasured heterogeneity.

This article was published Online First July 28, 2014.

Rebekah Levine Coley, Caitlin McPherran Lombardi, and Jacqueline Sims, Applied Developmental & Educational Psychology, Lynch School of Education, Boston College.

This article uses confidentialized unit record data from Growing Up in Australia: The Longitudinal Study of Australian Children. The study is conducted in partnership between the Australian Government Department of Families, Housing, Community Services and Indigenous Affairs (FaHCSIA), the Australian Institute of Family Studies (AIFS), and the Australian Bureau of Statistics (ABS). The findings and views reported in this article are those of the authors and should not be attributed to FaHCSIA, AIFS, or the ABS. This research was supported in part by a grant to the first author from the Spencer Foundation. The authors are grateful to Linda Harrison for her helpful insights into data from the Longitudinal Study of Australian Children Birth Cohort.

Correspondence concerning this article should be addressed to Rebekah Levine Coley, Applied Developmental & Educational Psychology, Boston College, Campion Hall 239A, 140 Commonwealth Avenue, Chestnut Hill, MA 02467. E-mail: coleyre@bc.edu

We base our analyses on theories of child development that point to the importance of responsive, stimulating caregiving and children's responses to environmental stressors in the first years of life. During infancy and early childhood, as children's cognitive, social, and emotional skills develop rapidly, supportive interactions, responsive caregiving, and safe, stimulating learning environments are especially important (Early & Burchinal, 2001). For infants in particular, a consistent and responsive child-caregiver relationship is essential to promote secure child-caregiver attachments while providing opportunities for children to safely explore their environment. Thus, nonparental care may be less supportive for infants' development, most notably when the care is provided with larger groups of children or less one-on-one child-adult interaction, as may be found in center programs (Dowsett, Huston, Imes, & Gennetian, 2008). Further, center EEC programs may expose young children to greater physiological stress, as recent research indicates that experiences in center EEC put infants and toddlers at greater risks than older children for cortisol elevations throughout the school day which in turn may impede children's early emotional and cognitive development (Dettling, Gunnar, & Donzella, 1999; Vermeer & van IJzendoorn, 2006; Watamura, Donzella, Alwin, & Gunnar, 2003). Preschool-aged children, in contrast, have developed enhanced language skills, emotional regulation, and social skills in comparison to their younger counterparts, and thus are likely to experience less stress from center-based EEC (Dettling et al., 1999; Vermeer & van IJzendoorn, 2006; Watamura et al., 2003). Center-based preschools may better support preschool-aged children's cognitive skills, as there are more opportunities to experience structured and diverse educational curricula in centers than in parent care or home-based EEC (Coley, Li-Grining, & Chase-Lansdale, 2006; Dowsett et al., 2008; Fuller, Kagan, Loeb, Chang, 2004; Maccoby & Lewis, 2003).

There are also theoretical perspectives to suggest that EEC experiences may be differentially influential across subgroups of children (Bradley, McKelvey, & Whiteside-Mansell, 2011). Children from low-resource home environments due to limited economic resources or low parental education typically experience less enriching, stimulating, warm, and consistent home environments than their counterparts in economically advantaged families (Magnuson & Votruba-Drzal, 2009). Compensatory models of EEC argue that EEC programs may provide a particularly important resource to bolster the early academic and behavioral skills of children from homes with limited socioeconomic resources (Loeb, Bridges, Bassok, Fuller, & Rumberger, 2007; Loeb, Fuller, Kagan, & Carroll, 2004; McCartney, Dearing, Taylor, & Bub, 2007; Votruba-Drzal, Coley, Koury, & Miller, 2013). Theories of differential susceptibility, in contrast, argue that EEC may be more influential for children with more difficult and challenging temperaments, although evidence supporting this theory are rather sparse and focus on EEC quality rather than type and quantity and on behavioral but not cognitive arenas of child functioning (Belsky & Pluess, 2011; Pluess & Belsky, 2010).

Empirical Review of Early Childhood Education and Children's School Readiness Skills

Understanding the repercussions of EEC programs for children's core cognitive and behavioral skills is essential, because such skills are key predictors of successful transitions into formal

schooling and long-term educational success (Entwistle & Alexander, 1993; Li-Grining, Votruba-Drzal, Maldonado-Carreño, & Haas, 2010). Below we briefly review the empirical evidence, drawn primarily from studies of American children.

Numerous studies have found that children who attend EEC, particularly center-based preschool programs, show enhanced early reading and numeracy skills in comparison to their peers in parent care or more informal home-based care settings, differences that extend into elementary school (Gormley, Gayer, Phillips, & Dawson, 2005; Loeb et al., 2007; Magnuson, Meyers, Ruhm, & Waldfogel, 2004; Morrissey, 2010; NICHD ECCRN, 2002, 2005; Duncan & NICHD ECCRN, 2003). One of the primary limitations of much of this work is the focus solely on preschool care. In nonexperimental studies that limit attention to children's EEC experiences in the year or two prior to formal school, models fail to adequately delineate whether associations between center-based preschool and children's enhanced cognitive skills in elementary school are driven by preschool experiences or rather by correlated experiences earlier in childhood. These models also fail to demarcate whether earlier EEC experiences show unique links with children's later cognitive skills.

A handful of studies from the United States suggest that center-based EEC during infancy and the early toddler years may be less beneficial for children's later cognitive skills than center EEC during the late toddler and preschool years (Loeb et al., 2007; Duncan & NICHD ECCRN, 2003; Votruba-Drzal et al., 2013). For example, Loeb et al.'s (2007) study using retrospective reports of the timing of center EEC initiation indicated starting center EEC between 2 and 3 years of age was associated with greater kindergarten cognitive skills than starting EEC earlier or later. Duncan and NICHD ECCRN (2003) found similar results using prospective data, arguing that greater exposure to center care between 2.5 and 4.5 years had the strongest associations with children's cognitive skills in kindergarten, with no benefits of earlier center EEC. One interpretation of these results rests on developmental timing, suggesting that center-based EEC is most cognitively supportive for children over age 2. A second interpretation is a duration of exposure argument, suggesting that 2 to 3 years of center care is more supportive of children's cognitive skill development than more or fewer years. A third argument concerns the intensity of exposure, suggesting that moderate rather than limited or extensive exposure to EEC is most beneficial (Votruba-Drzal et al., 2013). Few studies have tried to delineate the relative merits of these perspectives.

Only one published article of which we are aware has assessed such associations using nationally representative data from Australia (Coley, Lombardi, Sims, & Votruba-Drzal, 2013). Following children from infancy through the transition to elementary school, this article considered children's center EEC exposure at 9 months, 2 years, and 4 years and their cognitive, language, and reasoning skills at age 7. In contrast to a large body of research on American children, this study found no benefits of center-based preschool programs for 4-year-olds but, rather, reported that center care at age 2 was most predictive of children's later cognitive skills (Coley, Lombardi, et al., 2013). Although this study found benefits of both part-time and full-time center care, it did not assess the accumulation of EEC experiences over time, leaving open questions concerning whether

the apparent benefits of center EEC programs for toddlers were related to greater overall EEC exposure.

In contrast to the relatively consistent benefits of center-based EEC programs for children's cognitive skills, research regarding children's behavioral skills paints a less positive picture. For example, numerous longitudinal correlational studies find that children attending center-based EEC show higher rates of aggression, disruptiveness, and other externalizing problem behaviors than their peers who use informal home-based EEC or parent care, with small associations emerging during early childhood and remaining statistically significant through early elementary school and, in some cases, even into adolescence (Belsky et al., 2007; Magnuson et al., 2007). The strongest and most consistent associations appear to be with children's externalizing problems, with numerous studies also identifying links with lower attention skills and less consistent associations with prosocial behaviors, and with associations generally stronger with teacher than parent reports of children's behavioral functioning (Coley, Votruba-Drzal, et al., 2013; Loeb et al., 2007; Magnuson et al., 2007; NICHD ECCRN, 2003, 2006).

In the behavioral arena, previous research has unearthed evidence that both the duration and intensity of center EEC are associated with outcomes. For example, a number of studies have assessed children's accumulated months in center care or age at entry, finding that greater duration of center EEC predicted heightened behavior problems (Belsky et al., 2007; Loeb et al., 2007; NICHD ECCRN, 2003). Other studies have focused on the intensity of center exposure through hours per week, again finding a dosage effect linked to behavioral outcomes (Belsky et al., 2007; Coley, Votruba-Drzal, et al., 2013; Loeb et al., 2007; McCartney et al., 2010; NICHD ECCRN, 2003, 2006; Vandell, Belsky, Burchinal, Steinberg, & Vandergrift, 2010). Less evidence has emerged related to the developmental timing of center care and children's behavioral functioning, with a number of studies finding no significant patterns between the timing of EEC and children's later functioning (NICHD ECCRN, 2003; Peisner-Feinberg et al., 2001) and others finding that early entry into center care exacerbates negative associations with children's behavioral functioning in primary school (Coley, Votruba-Drzal, et al., 2013; Loeb et al., 2007).

Again, the majority of this evidence derives from studies of American children, with limited information concerning whether these patterns generalize into the Australian context. Greater amounts of center-based EEC in toddlerhood (2–3 years) have been associated with higher concurrent behavior problems among Australian children (Yamauchi & Leigh, 2011). Claessens and Chen (2013) found that center-based preschool was concurrently associated with higher prosocial skills and lower peer problems according to mother reports, with no significant effects for duration of care. In contrast, children experiencing multiple types of care showed lower prosocial skills and higher conduct problems. Australian studies have not assessed long-term associations between the duration and intensity of EEC and children's behavioral functioning following formal school entry or considered the relative role of early versus later EEC experiences.

An interpretational challenge of this broad range of correlational studies is the inability to determine whether there is a true causal relationship between EEC and children's later cognitive and behavior functioning, or rather whether selection factors may have

biased the measured associations. Studies using experimental or quasiexperimental techniques to assess the influence of state pre-kindergarten (pre-K; Gormley & Gayer, 2005; Gormley, Phillips, Newmark, Welti, & Adelstein, 2011) or Head Start programs (U.S. Department of Health and Human Services, Administration for Children and Families, 2010) have largely replicated cognitive skills benefits but have not found negative effects on behavioral functioning. There are four leading explanations for these different patterns. First, the experimental and quasiexperimental techniques may better adjust for selection bias, suggesting that the negative behavioral effects of preschool care in correlational research might be due to differential selection into EEC. Second, given the prevalence of EEC attendance in the United States, the control groups utilized in the experimental/quasiexperimental evaluations of pre-K and Head Start programs were composed of children in center care, Head Start, informal nonparental care, and parent care (Gormley et al., 2005, 2011; U.S. Department of Health and Human Services, Administration for Children and Families, 2010). This lack of a clean experimental manipulation of the treatment may have weakened the experimental effects. A third explanation is that many of the pre-K and Head Start programs assessed in experimental or quasiexperimental studies are held to more rigorous quality standards than most community-based EEC programs, suggesting that higher quality programs may limit negative behavioral effects on children (McCartney et al., 2010). Finally, these pre-K and Head Start programs served primarily lower income families. Some have argued that cognitive benefits of center EEC may be heightened for poor children, children of less educated parents, or children in families who provide less enriching home environments (Geoffroy et al., 2010; Loeb et al., 2004, 2007; McCartney et al., 2007; Votruba-Drzal et al., 2013), although other studies have found stronger detriments for behavioral functioning among poor children, as well (Loeb et al., 2007). Together these issues highlight the importance of more causal methodological techniques and of considering the moderating role of family socioeconomic status.

Early Childhood Education Policy in Australia

Although Australia and the United States share many cultural and economic similarities, the context of early childhood education and care shows some notable differences. These differences largely reflect the Australian government's broader and more generous supports to families with young children, as well as differences in family practices and maternal employment. For example, while the United States lacks a paid federal parental leave policy and has a limited unpaid federal leave policy, Australia has long offered a 1-year unpaid parental leave and a generous cash payment upon the birth of a child (recently expanded into a federal paid parental leave policy). Policy differences also exist for mothers receiving welfare in Australia versus the United States. Whereas poor mothers are imposed with work requirements early in their child's infancy in many states in the U.S., Australian mothers receiving welfare are not required to work until their youngest child is 6 years of age (Australian Government Department of Human Services, 2014). Together, these policies encourage more equitable access to resources and parental choice regarding employment and nonparental care for Australian families (Waldfogel, 2009) and may lead to lower use of nonparental care

for infants. Indeed, recent data suggest that 35% of Australian infants are in regular nonparental EEC at 9 months of age, in comparison to 50% of American children (Coley, Lombardi, et al., 2013).

The Australian government also provides greater financial support for families who do use EEC. In the United States direct government subsidies for EEC are reserved for low-income and poor families, with middle and upper income families receiving less generous tax credits. Further, most American families rely on the private market for EEC (U.S. General Accounting Office, 1997). In contrast, the Australian federal government covers up to half of families' center EEC costs (capped at \$7,500 yearly) through the Child Care Rebate and provides substantial subsidies for registered informal EEC including relative care (Australian Government Family Assistance Office, 2011; Michel, 2003). In addition, public preschools are funded directly by state governments in Australia, and programs are more heavily regulated. These differences translate into different prevalence rates: although preschool attendance is lower among economically and socially disadvantaged families in both countries (Dowling & O'Malley, 2009; Harrison & Ungerer, 2005), a higher proportion of 4-year-olds attend preschool in Australia, primarily part-time center preschool programs, whereas in the United States preschool attendance is lower but more likely to be full-time (Australian Bureau of Statistics, 2006; Coley, Lombardi, et al., 2013; Harrison & Ungerer, 2005; Harrison et al., 2009). For example, a recent comparison of nationally representative samples in the two countries found that 75% of Australian 4-year-olds attended center preschool programs, 63% part time and 11% full time at the time of the interview. In the United States, 69% of 4-year-olds were attending center-based preschools, equally split between part-time and full-time (Coley, Lombardi, et al., 2013). However, with the rapid expansion in publicly funded state pre-K programs in the United States that is currently underway, the Australian system provides an interesting example of what preschool options may look like in coming years.

In addition to these differences in financial support for EEC, public support and quality regulations for EEC differ between the countries as well. Like in the United States, quality regulations in Australia vary across states and territories, although the federal government has recently developed a National Quality Framework (NQF) regulation system that is replacing separate state licensing procedures. This system evaluates the majority of care settings in Australia with teacher education and care ratio requirements, which differs notably from the United States where roughly 25% of children experiencing care attend unregulated care settings (Zigler, Marsland, & Lord, 2009). These differing regulations result in varying experiences for children across the two countries. A recent comparison found that 96% of Australian 2-year-olds attending EEC centers were attending accredited programs, compared to only 32% in the United States; these differences were similar for preschoolers, with rates of accreditation among EEC centers attended by 4-year-olds at 100% in Australia and only 49% in the United States. Similar differences emerged in teacher training, with 82% of Australian 2-year-olds but only 23% of American 2-year-olds in center EEC programs having a head teacher with a degree in early childhood education or a related field (Coley, Lombardi, et al., 2013).

In short, policy comparisons suggest that the more generous and flexible family policies allow Australian parents more freedom than their American counterparts in selecting and affording EEC for their young children, which may lead to more stable EEC experiences and greater parental satisfaction with EEC choices. Greater EEC regulations similarly may lead to higher quality EEC in Australia than the U.S., although this supposition has not been directly assessed. These differences in the EEC context lead to the hypothesis that the benefits of EEC programs for children's cognitive skills may be stronger in Australia, and perhaps that detriments in terms of children's behavior problems may be lessened.

Present Study

In summary, research findings on EEC have been robust across numerous longitudinal data sets of children from the United States; however, little research has sought to replicate this research in other countries to assess the generalizability and universality of findings. With a similar economic and social structure and yet some differences in EEC use, funding, and quality controls, Australia offers a context to replicate this research while also examining policy differences. The primary goals of this research were to delineate the contributions of the duration, intensity, and timing of exposure to center-based EEC to children's cognitive skills and behavioral functioning following entry to formal schooling, utilizing a nationally representative sample of Australian children followed prospectively from infancy through first grade and rigorous statistical methods to help adjust for differential selection into EEC. Based upon past research conducted mostly in the United States, we expected that greater exposure to center EEC through an earlier age of entry or greater hours per week would be associated with both higher cognitive skills as well as higher behavior problems for children. In terms of the timing of EEC, we expected that center care after age 2 would be more positively linked to cognitive skills than infant center care, whereas for behavioral functioning, we expected that a greater duration or intensity of exposure to center care would be associated with greater behavior problems and lower attention skills, with less consistent associations with prosocial behaviors. Given the greater flexibility and perhaps higher quality of care in Australia, we expected that cognitive effects would be somewhat stronger and behavioral effects somewhat weaker than results that U.S. studies have unearthed.

Method

Sample

Data for this article were drawn from the Longitudinal Study of Australian Children Birth Cohort (LSAC), a multimethod study seeking to document children's development and proximal environments from infancy through childhood. The LSAC sampled a nationally representative cohort of 5,107 children born in Australia between March 2003 and February 2004. Births were sampled from the Medicare enrollment database, in which all Australian children are enrolled. Stratification was used to ensure proportional geographic representation for each state and territory. The survey sample excluded nonpermanent residents and children with the same name as deceased children, and only allowed for one child per household. A study comparing the LSAC population with

the 2001 Census population found that the LSAC largely mirrored the general population with few differences across a wide range of demographic measures (Soloff, Lawrence, Misson, & Johnstone, 2006). For more information on the LSAC, see Soloff, Lawrence, and Johnstone (2005).

The LSAC has collected seven waves of data with in-person interviews and direct assessments as well as mail-in surveys. In-person interviews with parents occurred when children were 9 months (Wave 1), 3 years (Wave 2), 5 years (Wave 3), and 7 years (Wave 4) with response rates of 58%,¹ 90%, 86%, and 84%, respectively. Mail-in written parent surveys were collected every in-between year when children were 2 years (Wave 1.5), 4 years (Wave 2.5), and 6 years (Wave 3.5) with response rates of 70%, 64%, and 59%, respectively. At each wave there was some variability in the age at which children were assessed, with standard deviations averaging about 3 months at each wave (see Table 1) and approximately 90% of children falling within ± 4 months of the target age at each assessment. Age at assessment was included as a covariate in all models to help adjust for these differences.

The analytic sample included all children in the Wave 1 sample, $N = 5,107$. Within the analytic sample there were missing observations due to attrition over the waves and missing data on individual measures. Because missing data introduce biases into the sample, missing data were imputed using multiple imputation by chained equations, implemented in Stata 12 (Royston, 2004, 2005) to create 10 complete data sets. After imputation, survey weights, which adjust for selection criteria and differential response, were incorporated in all analyses. The use of these weights makes the sample representative of infants born in Australia between March 2003 and February 2004.

The LSAC offers a number of strengths for the purposes of this research. The data are nationally representative and hence present a generalizable sample of young children. The sample is large and includes sizable subsamples of economically disadvantaged children and Aboriginal children. Moreover, the sample contains very strong measurements of children's development using reliable and well-validated instruments. Data on child functioning were collected from three different sources at age 7: direct assessments, parent reports, and teacher reports. The use of multiple sources of information on children's functioning is important on many fronts. First, it helps to assuage analytic concerns over shared method variance. Second, parent and teacher reports of children's behavioral functioning and direct assessments and teacher reports of children's cognitive skills are only moderately correlated (due both to differences in children's functioning across contexts and to differences in raters' expectations and impressions of children; Cabell, Justice, Zucker, & Kilday, 2009; Kilday, Kinzie, Mashburn, & Whittaker, 2012; Strickland, Hopkins, & Keenan, 2012), yet all show predictive validity to longer-term functioning suggesting that they provide unique windows into children's well-being.

Another strength of the LSAC is that it provides five points of data from infancy to preschool, providing rich information on children's EEC experiences through early childhood, and followed children through school entry. At the same time, it is important to acknowledge limitations, namely, that data on EEC were only collected at discrete time points and do not include a full account of EEC experiences from birth through preschool and that the developmental quality of care, another important characteristic of

early care experiences, was not assessed in the LSAC and hence could not be taken into account.

Measures

EEC characteristics. At Waves 1 through 3 (Waves 1, 1.5, 2, 2.5, and 3) parents reported on children's regular nonparental care settings. At each wave, EEC type was coded into three mutually exclusive categories designating center-based (day care center, preschool, or other center-based child care program), informal care (grandparent, other relative, nanny, other nonrelative, family day care, occasional care, gym/leisure/community center, mobile care unit), or parent care. Children in regular nonparental care settings for less than 5 hr per week and children who did not experience any nonparental care were coded as being in parent care. Children in center care, including those attending only center care and those attending both center care and informal care, were coded as being in center care.² Children only in informal care were coded as informal care. Mothers also reported the total number of hours per week that children spent in EEC at Waves 1, 1.5, 2, 2.5, and 3 of the survey, coded in units of 10. At Wave 3, about 20% of children had entered kindergarten; for these children only four waves of data on EEC were available. For the 80% of children who had not entered kindergarten by Wave 3, five waves of data on EEC were assessed.

These EEC measures were then used to create three sets of variables to delineate children's duration, intensity, and timing of exposure to center-based EEC. The duration of center EEC was measured through a variable indicating the percentage of waves in center care. Another variable delineated the percentage of waves in informal care. The intensity of center EEC was measured with the average number of hours children spent in center EEC (hours were coded as 0 for waves in which children were not in any center care). Finally, to assess the timing of center EEC, we created six mutually exclusive categories: (a) parent or informal care only from infancy until primary school, (b) center care from infancy (Wave 1) through preschool (Waves 2.5 and/or 3), (c) center care from toddlerhood (Waves 1.5 and/or 2) through preschool, (d) center care only during preschool (Waves 2.5 and/or 3), (e) center care in infancy and/or toddlerhood but not in preschool, and (f) inconsistent center care that included some center care in infancy and/or toddlerhood plus preschool. Initial analyses of these groups determined that there were no differences in functioning between the children who entered center care in infancy and stayed (Group 2), entered center care in toddlerhood and stayed (Group 3), and children who entered center care early and attended preschool but spent at least one wave in parent or informal care (Group 6). Thus we collapsed these three categories into one, resulting in four mutually exclusive categories indicating the timing of center care: (a) no center care from infancy through preschool, (b) center care in infancy and/or toddlerhood and preschool, (c) center care only

¹ Different response rates have been reported based on different calculations. This response rate includes nonresponse from all sources from the originally drawn sample (see Gray & Sanson, 2005).

² Center care was prioritized in this manner due to extant literature suggesting the significant role of center care in children's development (Loeb, Bridges, Bassok, Fuller, & Rumberger, 2007; Magnuson, Meyers, Ruhm, & Waldfogel, 2004; Morrissey, 2010).

Table 1
Sample Descriptives

Variable	Wave 1			Wave 1.5			Wave 2			Wave 2.5			Wave 3			Wave 4		
	<i>M</i>	<i>SD</i>	%	<i>M</i>	<i>SD</i>	%												
Care type																		
Parent			65.46			47.39			33.17			14.53			6.98			
Center			10.69			26.16			41.61			74.69			90.64			
Informal			23.85			26.49			25.22			10.78			2.38			
Care duration																		
% of waves center care	0.46	0.26																
% of waves informal care	0.16	0.21																
Care intensity																		
Hours of care/10	0.23	0.78		0.51	1.06		0.84	1.28		1.34	1.22		1.59	1.46				
Average hours of care/10	0.94	0.81																
Timing of care																		
Parent only care			7.63															
Early center plus preschool			44.87															
Preschool only			45.27															
Early center, no preschool			2.23															
Child outcomes																		
Academic skills															3.32	0.75		
Matrix reasoning															10.60	3.02		
Vocabulary															74.00	5.15		
Parent Attention Skills															0.89	0.45		
Parent conduct problems															0.32	0.29		
Parent prosocial skills															1.66	0.34		
Teacher attention skills															0.98	0.51		
Teacher conduct problems															0.20	0.28		
Teacher prosocial skills															1.50	0.44		
Covariates																		
Multiple care arrangements			8.41			17.96			18.54			36.65			39.83			
Child age	8.86	2.57		21.49	3.38		34.04	2.92		47.28	3.34		57.74	2.86		81.98	3.15	
Child male			51.16															
Child low birth weight			6.02															
Child bad health			3.12															
Child temperament	4.45	0.62																
Child cognitive skills	25.88	9.70																
Parent Asian			8.51															
Parent Aboriginal			4.58															
Immigrant household			31.50															
Non-English household			15.67															
Child number siblings	0.99	1.07					1.28	1.03					1.51	1.01				
Parent married			73.34						71.42							73.09		
Youngest parent's age	30.41	5.29																
Parent < high school education			6.90							6.01						5.37		
Parent high school education			6.50						5.67							4.69		
Parent some college			48.01						49.27							49.31		
Parent college/grad school			38.59						39.05							40.63		
Household income/10,000	6.89	4.16					7.83	4.86					9.62	5.53				
Low income			19.94															
Mother employed			32.51						49.83							58.69		
Cognitive stimulation							1.92	0.56							1.65	0.57		

during preschool, and (d) center care in infancy and/or toddlerhood but not in preschool.

In addition to the main EEC variables of interest, indicator variables were created across waves denoting whether children were in multiple EEC arrangements, given recent research suggesting that multiple care arrangements may be linked to worse functioning in children (Claessens & Chen, 2013; Morrissey, 2009). These variables assessed whether children experienced

multiple care arrangements at all waves (Waves 1, 1.5, 2, 2.5, and 3), some waves, or did not experience multiple care experiences at any wave (reference).

Children's outcomes. Measures of child functioning were drawn from Wave 4, when children averaged 7 years of age and were typically in Year 1 (first grade) of primary school. Wave 4 was chosen because child functioning measures were not assessed at Wave 3.5, and at Wave 3 about 80% of children had not yet

entered primary school and were still attending EEC programs. Three measures of children's cognitive skills were assessed at age 7 using direct assessments and teacher reports: vocabulary, academic skills, and matrix reasoning. Children's receptive vocabulary skills were directly assessed by field interviewers using a shortened version of the Peabody Picture Vocabulary Test (3rd ed.; PPVT-III; Dunn & Dunn, 1997). The PPVT was scored using item response theory (IRT) and then transformed to generate a scale with a mean of 64 and a standard deviation of 8. Children's academic skills were assessed with teacher reports using the Language and Literacy and Mathematical Thinking subscales from the Academic Rating Scale (ARS; National Center for Educational Statistics, 2002). The Language and Literacy Scale ($\alpha = .96$) had nine items (e.g., conveys ideas when speaking, reads fluently) that rate a child's performance in oral and written language according to a 5-point scale (*not yet* = 1, *beginning* = 2, *in progress* = 3, *intermediate* = 4, and *proficient* = 5). The Mathematical Thinking Scale ($\alpha = .94$) used the same scale to rate a child's performance on nine spatial and math items (e.g., creates and extends patterns, recognizes shape properties and relationships). Due to the high correlation between the two scores ($r = .81$), the measures were averaged to create one composite assessing teacher-reported language, literacy and mathematical thinking, termed "academic skills." The final measure of children's cognitive achievement was measured with the Matrix Reasoning (MR) test from the Wechsler Intelligence Scale for Children (4th ed.; WISC-IV). This test of nonverbal and fluid intelligence (35 items) presents the child with an incomplete set of diagrams and requires the child to select the picture that completes the set from five different options.

Children's behavioral functioning was reported separately by parents and by teachers using items from the Strengths and Difficulties Questionnaire (SDQ; Goodman, 1997), which rates children's skills on a 3-point scale (*not true* = 0, *somewhat true* = 1, and *certainly true* = 2). Factor analyses run separately by reporter derived three subscales assessing attention skills, prosocial behaviors, and conduct problems. For each subscale, items were averaged into a total score, leading to three scores from parent reports and three scores from teacher reports. The attention skills subscales ($\alpha_p = .78$; $\alpha_t = .88$) each included five items assessing children's ability to sit still, fidgeting, distractibility, thinking before acting, and attention span. Higher scores indicate greater attention and learning behaviors. The prosocial behaviors subscales ($\alpha_p = .70$; $\alpha_t = .83$) were composed of five items assessing children's considerate, sharing, helpful, kind, and volunteering behaviors. Also composed of five items, the conduct problems subscales ($\alpha_p = .60$; $\alpha_t = .76$) covered children's temper tantrums, obedience, fighting, lying or cheating, and stealing behaviors.

Child characteristics. A broad range of child and family characteristics were assessed through parent report. Child characteristics included age of assessment (in months) and gender. Child low birthweight status was represented with an indicator of whether the child was born low (less than 2,500 grams) birthweight. Child health condition was also represented by an indicator which reflected whether the child was of fair or poor health based on parent-report at Wave 1. Cognitive ability was assessed at Wave 1 using the Communication and Symbolic Behavior Scales Developmental Profile: Infant-Toddler Checklist (CSBS DP; Wetherby & Prizant, 2001). The CSBS (24 items, $\alpha = .89$) yields a standardized normed score of children's early social, language

and cognitive skills. Child temperament was measured with a shortened version of the Australian revision of the Toddler Temperament Scale (TTS; Fullard, McDevitt, & Carey, 1984), with four items rated on 6-point scales assessing children's abilities in each of three domains: approach, persistence, and reactivity ($\alpha = 0.98-0.99$). These three domains were combined into a composite measure with higher scores indicating an easier temperament with more approachability, persistence, and regulation.

Parental and household characteristics. Several parental and household characteristics were also assessed. Time-varying characteristics were measured at Waves 1, 2, and 3 and were coded to tap into shifts over time. Parent race/ethnicity was indicated with two dummy variables indicating (a) Asian origin and (b) Aboriginal origin of either parent. A dichotomous variable indicated whether either parent was an immigrant to Australia. An additional dichotomous variable indicated whether the primary language of the household was non-English. Family structure covariates included maternal marital status, measured categorically, delineating whether the respondent was consistently married (vs. single or cohabiting) at all waves, married at some waves, or not married at any wave (reference), and the number of children under age 18 in the household, measured at Wave 1 and then measured as changes in the number of children at Waves 2 and 3. Parental age was measured with a continuous measure of the age in years of the youngest parent in the household at Wave 1. Parental education was assessed using the highest level of educational attainment that parents reported at Waves 1 through 3 of data collection (shifts over time were not assessed due to limited change). Categorical indicators designated parents who had earned less than a high school degree, a high school degree but no college (reference), above a high school degree but less than a college degree, and a Bachelor's degree or higher. Total household annual income was expressed in units of 10,000 at Wave 1 and then measured as changes in income at Waves 2 and 3. Maternal employment was measured categorically across waves, delineating whether mothers were consistently employed across all waves, employed at some waves, or not employed at any wave (reference). At Waves 2 and 3 of the survey, parents' provision of enriching home environments was assessed. Parents reported the weekly frequency of a variety of activities such as drawing pictures with, reading to, and playing outdoors with their child (seven items). Responses ranged from 0 (*none*) to 3 (*6-7 days per week*) and were averaged within each wave, creating a measure of cognitive stimulation at Wave 2 and changes in stimulation at Wave 3 ($\alpha = 0.70-0.71$).

Analytic Approach

The primary goal of the analyses was to assess how exposure to center EEC from infancy through preschool was associated with Australian children's cognitive and behavioral skills following school entry. This question was assessed using a series of ordinary least squares (OLS) regression models predicting children's functioning at Wave 4 from their duration, intensity, and timing of center EEC exposure. Given that prior research has identified notable differences between characteristics of children in parental or informal care and children in center EEC programs (Coley, Votruba-Drzal, Collins, & Miller, 2014; Meyers & Jordan, 2006), a primary concern for the present study is that selection processes

rather than child care experiences per se may explain any associations with children's cognitive and behavioral functioning. To address this significant concern, three techniques were incorporated in the analyses. First, all OLS regression models incorporated a large set of child, maternal, and household characteristics drawn from Waves 1 through 3 as covariates. Just as the EEC variables were aggregated over Waves 1 through 3 to capture children's exposure to EEC throughout early childhood, the covariates similarly were coded to assess children's changing contexts over Waves 1 through 3. Covariates were selected (listed in Table 1) because they have been shown to be associated with selection into child care in prior research (Coley et al., 2014; Meyers & Jordan, 2006), although even the most thorough set of covariates leaves open the potential for omitted variable bias (Duncan, Magnuson, & Ludwig, 2004). As a second mechanism helping to control for unmeasured variable bias, models were run as lagged regressions, incorporating a Wave 1 measure of cognitive ability (for models predicting cognitive skills) or a Wave 1 measure of child temperament (for models predicting behavioral outcomes) as an additional covariate to control for unmeasured, time-invariant factors that have a consistent effect on children's functioning (Cain, 1975), thus further reducing concerns of omitted variable bias.

Third, propensity score weighting (PSW) techniques were used to help further adjust for potential selection bias (Imbens, 2000; Rosenbaum & Rubin, 1983). Propensity score (PS) techniques help to equate respondents on observed, preexisting characteristics (Rosenbaum & Rubin, 1983). Propensity score techniques have been shown to remove as much as 90% of selection bias in nonexperimental research (Leon & Hedeker, 2007), although it is important to note that PS techniques cannot control for unobserved factors, the influence of which may even be magnified by matching on observables (Pearl, 2009). We incorporated propensity scores using the three step weighting procedure described by Imbens (2000). This propensity score weighting procedure is highly flexible as it is able to accommodate both continuous and categorical measures such as our primary measures of center EEC (Imbens, 2000). The first step involved estimating each child's propensity to receive the "treatment," that is to be in center EEC. For the continuous measures of duration (% waves in center EEC) and intensity (average hours of center EEC), the first step used OLS regression models to estimate the propensity of having a higher % of waves or more hours in center care as a function of observed pretreatment covariates (all child and family characteristics noted above drawn from Wave 1). For the categorical timing of center EEC variable, multinomial logistic regression models were used to estimate the propensity of being in each of the four patterns of EEC as a function of all observed pretreatment (Wave 1) covariates. Appendix Table A1 presents results of these models using one randomly selected imputed data set as an exemplar. In the second step, propensity score weights were created by taking the inverse of each child's conditional probability of receiving the EEC treatment that the child actually received (Imbens, 2000). In the third step, we incorporated the propensity score weights in the lagged longitudinal regression models predicting child functioning at age 7. These models were run weighted with the EEC treatment-specific propensity score weights multiplied by the sample weights to generate the average treatment effect of the EEC experience, as shown in Equation 1.

$$\begin{aligned} \text{Child Outcome}_{4i} = & B_0 + B_1 \text{EEC}_{1-3i} + B_2 \text{Child Outcome}_{1i} \\ & + B_3 \text{Maternal}_{1-3i} + B_4 \text{Child}_{1-3i} + \epsilon_i. \end{aligned} \quad (1)$$

After the first set of models assessing associations between children's duration, intensity, and timing of exposure to center EEC and their cognitive and behavioral skills, a second set of analyses were estimated to address whether associations between EEC experiences and children's later functioning differed depending on family income, parental education, home environment, or child temperament. Income and home environment moderation were tested with interactions between centered, continuous measures of income (averaged over Waves 1-3) and home environments (averaged over Waves 2 and 3) and each of the EEC variables. Education moderation was tested with interactions between categorical indicators of parental education (less than high school, some college, and a college or graduate degree) and each of the EEC variables, with posthoc analyses to address differences between groups. Moderation by child temperament was assessed with interactions between children's centered Wave 1 temperament measure and each of the EEC variables.

Results

Sample Descriptives

Descriptive statistics for the EEC measures, child and family covariates, and children's age 7 functioning are displayed in Table 1. Participation in EEC grew from 35% during infancy to 93% when children were nearly 4 years old (among children not yet in kindergarten), with notable growth in center care and decline in parent and informal care. Overall, children were in center-based programs nearly half of the time from infancy through age 4, spending an average of 46% of waves in center care and 16% of waves in informal care. The intensity (hours per week) of center EEC also increased, although by the preschool-age wave children averaged only 16 hr per week in center care. Over all time periods, children experienced an average of less than 10 hr per week in center care (8% had 0 hr, 59% less than 10, 24% averaged 10 to 20 hr, and only 10% of children averaged 20 hr or more per week of center care).

Although the patterns of increasing use and intensity of center EEC suggests a linear increase, an examination of individual children's experiences found several distinct patterns of center EEC exposure for Australian children. Approximately 45% of children were in center care at all waves or nearly all waves, including both during infancy or toddlerhood and during preschool. Another 45% of children only used center care during preschool. A very small percentage (2%) attended center care at some point during infancy or toddlerhood but did not attend preschool. Another small group (8%) did not attend center care at all prior to entering kindergarten.

EEC Predicting Child Cognitive Skills at Age 7

The first set of models shown in the top panel of Table 2 (Model 1) examined whether there were differences in children's cognitive skills at age 7 depending on the duration of center and informal

Table 2
Propensity Score Weighted OLS Regression Models With Duration, Intensity, and Patterns of EEC Predicting Child Cognitive Skills at Age 7

Independent variable	Teacher academic skills	Matrix reasoning	Vocabulary
Model 1: Duration of care			
% waves center care	0.13 (0.07) [†]	0.65 (0.23)**	0.12 (0.41)
% waves informal care	0.05 (0.08)	0.10 (0.28)	-0.21 (0.52)
Covariates			
Multiple care all waves	0.01 (0.10)	-0.38 (0.44)	0.73 (0.72)
Multiple care some waves	-0.01 (0.04)	-0.01 (0.14)	0.13 (0.26)
Child age	0.03 (0.01)**	-0.04 (0.02)*	0.24 (0.03)**
Child male	-0.12 (0.03)**	-0.09 (0.09)	0.51 (0.17)**
Child low birth weight	-0.21 (0.06)**	-0.39 (0.24)	-0.16 (0.42)
Child bad health	0.00 (0.07)	0.10 (0.30)	-0.88 (0.46) [†]
Child cognitive skills	0.01 (0.00)*	0.01 (0.01)	0.02 (0.01)*
Parent Asian	0.09 (0.07)	0.38 (0.29)	-0.50 (0.44)
Parent Aboriginal	-0.15 (0.08) [†]	-0.02 (0.3)	-0.94 (0.53) [†]
Immigrant household	0.00 (0.03)	0.24 (0.13) [†]	0.02 (0.22)
Non-English household	-0.02 (0.06)	-0.18 (0.21)	-1.80 (0.33)**
Wave 1 siblings	-0.06 (0.02)**	-0.20 (0.06)**	-0.69 (0.11)**
Wave 2 change in siblings	0.05 (0.03) [†]	-0.01 (0.11)	-0.16 (0.18)
Wave 3 change in siblings	0.02 (0.03)	0.07 (0.11)	0.18 (0.19)
Parent married all waves	0.16 (0.04)**	0.21 (0.16)	0.35 (0.25)
Parent married some waves	0.09 (0.05) [†]	0.05 (0.20)	0.18 (0.30)
Youngest parent's age	0.01 (0.00)*	0.03 (0.01)**	0.14 (0.02)**
Parent < high school education	-0.19 (0.09) [†]	-0.52 (0.40)	-0.53 (0.64)
Parent some college	-0.02 (0.06)	0.06 (0.24)	0.63 (0.40)
Parent college/grad school	0.13 (0.06)*	0.65 (0.25)*	1.66 (0.41)**
Wave 1 household income	0.14 (0.03)**	0.49 (0.14)**	0.87 (0.24)**
Wave 2 change in income	0.06 (0.04)	0.63 (0.17)**	0.47 (0.28)
Wave 3 change in income	0.08 (0.04)*	0.09 (0.18)	-0.06 (0.25)
Mother employed all waves	0.04 (0.04)	0.08 (0.16)	0.16 (0.29)
Mother employed some waves	0.04 (0.03)	-0.03 (0.13)	0.26 (0.22)
Wave 2 Cognitive stimulation	0.05 (0.03)	0.20 (0.12)	1.12 (0.18)**
Wave 3 Change in cognitive stimulation	-0.01 (0.03)	-0.05 (0.11)	0.31 (0.20)
F of model	15.28	7.74	21.05
R ²	0.15	0.07	0.19
Model 2: Intensity of care			
Avg. hours in center care	0.02 (0.07)	0.61 (0.29)*	0.39 (0.48)
F of model	13.09	7.63	18.01
R ²	0.14	0.07	0.19
Model 3: Patterns of care			
Early center plus preschool	0.02 (0.07)	0.61 (0.29)* ^a	0.39 (0.48)
Preschool only	-0.03 (0.07)	0.29 (0.27) ^a	0.09 (0.44)
Early center, no preschool	-0.13 (0.15)	0.90 (0.49)	-0.13 (0.79)
F of model	4.44	2.35	5.56
R ²	0.16	0.13	0.19

Note. OLS = ordinary least squares; EEC = early education and care; Avg. = average. Models 2 and 3 included all covariates. Within columns, matched superscripts indicate difference at $p < .05$.

[†] $p < .10$. * $p < .05$. ** $p < .01$.

EEC that children experienced from infancy until kindergarten entry. Models were weighted with PSW weights that adjusted for each child's propensity of being in higher percentages of center care, based upon observed characteristics from Wave 1. Due to limitations of PSW, we could not adjust for both the percentage of center EEC and percentage of informal EEC within the same model, however we ran models adjusting for each and found the results to be nearly identical between sets of models, so present only the models adjusting for the percentage of center EEC. Results from Model 1 indicate that greater exposure to center care

from infancy through preschool was associated with enhanced cognitive skills at age 7. The effect sizes were very small. A one-standard-deviation (SD) difference in waves of center care was predictive of .06 SDs higher matrix reasoning scores with nonsignificant results for academic and vocabulary skills. The percentage of waves in informal care was not significantly associated with any measure of children's cognitive skills, supporting the primacy of center EEC programs in children's functioning. Similarly, exposure to concurrent multiple care arrangements was not predictive of children's cognitive skills in first grade.

The second panel in Table 2 (Model 2) considered children's intensity of exposure to center EEC (the hours per week children attended centers from infancy through preschool). Greater intensity of center EEC was associated with higher matrix reasoning skills, with a small effect size of .16 *SDs* but was not linked to children's academic or vocabulary skills. Results from the third set of models assessing the role of center EEC timing are presented in the third panel (Model 3). These models found that children in center care during both infant/toddler as well as preschool years had higher matrix reasoning skills after school entry than did their peers who only attended center-based preschool (.11 *SD*, indicated

by shared superscripts) or who were consistently in parent care (.20 *SD*). The timing of care was not significantly related to children's academic or vocabulary skills.

EEC Predicting Child Behavioral Skills at Age 7

A parallel set of models was run predicting both parent reports and teacher reports of children's behavioral skills at age 7, with results presented in Table 3. From the first set of models examining duration of center and informal EEC, results found that greater exposure to center care from infancy through preschool was asso-

Table 3
Propensity Score Weighted OLS Regression Models With Extent, Intensity, and Patterns of EEC Predicting Child Behavioral Functioning at Age 7

Independent variable	Parent attention	Parent conduct	Parent prosocial	Teacher attention	Teacher conduct	Teacher prosocial
Model 1: Duration of care						
% waves center care	-0.10 (0.04)*	0.04 (0.02)†	-0.05 (0.03)	-0.07 (0.05)	0.06 (0.03)*	-0.02 (0.04)
% waves informal care	-0.09 (0.04)*	0.04 (0.03)	-0.03 (0.04)	-0.07 (0.05)	0.03 (0.03)	-0.02 (0.05)
Covariates						
Multiple care all waves	-0.07 (0.07)	-0.02 (0.04)	0.04 (0.05)	-0.03 (0.07)	-0.03 (0.05)	0.04 (0.07)
Multiple care some waves	-0.02 (0.02)	0.03 (0.01)*	-0.01 (0.02)	-0.04 (0.02)	0.01 (0.01)	-0.03 (0.02)
Child age	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
Child male	-0.18 (0.02)**	0.07 (0.01)**	-0.15 (0.01)**	-0.32 (0.02)**	0.09 (0.01)**	-0.24 (0.02)**
Child low birth weight	-0.06 (0.04)	0.02 (0.02)	-0.01 (0.03)	-0.06 (0.04)	0.00 (0.02)	-0.01 (0.04)
Child bad health	-0.11 (0.05)*	0.03 (0.03)	-0.05 (0.03)	-0.02 (0.06)	0.02 (0.03)	-0.05 (0.05)
Child temperament	0.05 (0.01)**	-0.05 (0.01)**	0.05 (0.01)**	-0.04 (0.01)**	0.02 (0.01)†	-0.02 (0.01)
Parent Asian	0.04 (0.04)	-0.05 (0.02)†	-0.01 (0.03)	0.09 (0.04)*	-0.03 (0.02)	-0.03 (0.04)
Parent Aboriginal	-0.02 (0.05)	0.02 (0.03)	-0.03 (0.03)	-0.14 (0.06)*	0.10 (0.04)**	-0.12 (0.06)*
Immigrant household	0.02 (0.02)	-0.03 (0.01)*	-0.02 (0.01)	-0.01 (0.02)	0.02 (0.01)	-0.02 (0.02)
Non-English household	-0.02 (0.03)	0.04 (0.02)*	-0.02 (0.02)	-0.05 (0.03)	0.01 (0.02)	-0.03 (0.03)
Wave 1 siblings	0.01 (0.01)	0.01 (0.01)	-0.02 (0.01)*	-0.01 (0.01)	0.01 (0.01)	0.00 (0.01)
Wave 2 change in siblings	0.05 (0.02)**	0.00 (0.01)	-0.01 (0.01)	0.08 (0.02)**	-0.04 (0.01)**	0.05 (0.02)**
Wave 3 change in siblings	0.01 (0.02)	0.03 (0.01)**	-0.02 (0.01)	0.04 (0.02)†	-0.01 (0.01)	0.02 (0.02)
Parent married all waves	0.07 (0.02)**	-0.05 (0.01)**	0.02 (0.02)	0.08 (0.03)*	-0.05 (0.02)*	0.05 (0.02)†
Parent married some waves	0.03 (0.02)	-0.01 (0.02)	0.00 (0.02)	0.04 (0.04)	-0.02 (0.03)	0.05 (0.03)†
Youngest parent's age	0.00 (0.00)†	0.00 (0.00)	0.00 (0.00)	0.01 (0.00)**	0.00 (0.00)*	0.00 (0.00)*
Parent < high school	-0.04 (0.06)	0.04 (0.04)	-0.03 (0.04)	-0.03 (0.07)	-0.02 (0.04)	-0.01 (0.05)
Parent some college	-0.02 (0.04)	0.00 (0.03)	0.03 (0.03)	-0.02 (0.06)	0.00 (0.03)	-0.05 (0.04)
Parent college/grad school	0.03 (0.04)	-0.03 (0.03)	0.01 (0.03)	0.01 (0.06)	-0.01 (0.03)	-0.06 (0.05)
Wave 1 household income	0.10 (0.02)**	-0.05 (0.02)**	0.03 (0.02)	0.08 (0.02)**	-0.03 (0.02)†	0.07 (0.02)**
Wave 2 change in income	0.07 (0.02)**	-0.04 (0.02)*	0.01 (0.02)	0.08 (0.03)*	-0.04 (0.02)†	0.03 (0.02)
Wave 3 change in income	0.01 (0.03)	-0.01 (0.02)	0.00 (0.02)	-0.02 (0.03)	0.00 (0.01)	0.00 (0.02)
Mother employed all waves	0.04 (0.02)	-0.04 (0.02)*	0.02 (0.02)	0.03 (0.03)	-0.02 (0.02)	0.01 (0.03)
Mother employed some waves	0.02 (0.02)	-0.03 (0.01)*	0.03 (0.02)	0.05 (0.02)*	-0.02 (0.01)†	0.03 (0.02)
Wave 2 Cognitive stimulation	0.05 (0.02)**	-0.05 (0.01)**	0.09 (0.01)**	0.04 (0.02)†	-0.02 (0.01)	0.03 (0.02)†
Wave 3 Change in cognitive stim.	0.02 (0.02)	-0.03 (0.01)*	0.06 (0.02)**	-0.01 (0.02)	0.00 (0.01)	0.00 (0.02)
<i>F</i> of model	10.93	8.21	12.98	17.21	5.46	10.33
<i>R</i> ²	0.09	0.08	0.10	0.15	0.07	0.10
Model 2: Intensity of care						
Avg. hours in center care	-0.05 (0.01)**	0.02 (0.01)**	-0.02 (0.01)†	-0.05 (0.01)**	0.04 (0.01)**	-0.03 (0.01)*
<i>F</i> of model	10.23	7.79	12.16	16.31	5.52	9.10
<i>R</i> ²	0.10	0.08	0.11	0.16	0.08	0.11
Model 3: Patterns of care						
Early center plus preschool	-0.09 (0.05)	0.03 (0.03)	-0.02 (0.05)	-0.04 (0.05)	0.01 (0.03) ^a	-0.01 (0.05)
Preschool only	-0.06 (0.04)	0.01 (0.03)	0.01 (0.04)	-0.01 (0.05)	-0.01 (0.02) ^a	0.01 (0.05)
Early center, no preschool	-0.11 (0.07)	0.07 (0.05)	-0.04 (0.05)	-0.04 (0.08)	-0.02 (0.04)	-0.04 (0.08)
<i>F</i> of model	2.83	1.90	2.67	4.08	2.25	2.33
<i>R</i> ²	0.13	0.09	0.12	0.16	0.08	0.11

Note. Models 2 and 3 included all covariates. Within rows matched superscripts indicate difference at $p < .05$. OLS = ordinary least squares; EEC = early education and care; Avg. = average.

^a $p < .10$. * $p < .05$. ** $p < .01$.

ciated with lower parent-reported attention skills (.06 *SD*) as well as higher teacher-reported conduct problems (.06 *SD*) at age 7. Like with children's cognitive skills, neither the percentage of waves in informal EEC nor children's exposure to multiple concurrent care arrangements were significantly associated with children's later behavioral functioning with one exception: greater exposure to informal care predicted lower parent-reported attention skills (.04 *SD*).

Results for the intensity of center EEC shown in the following panel found a much more consistent pattern in which greater hours spent in center care were associated with lower behavioral functioning across five of the six outcome measures, again with small effect sizes. A one-*SD* increment in average hours of center care predicted increased levels of both parent- and teacher-reported conduct problems (.06 *SD* and .12 *SD*); decreased parent- and teacher-reported attention skills (.09 *SD* and .08 *SD*); and decreased teacher-reported prosocial behaviors (.06 *SD*).

In relation to the timing of care, only one significant effect emerged, showing that children in center care during both infant/toddler as well as preschool years had higher teacher-reported conduct problems after school entry than did their peers who only attended center EEC in preschool, with a small effect size of .07 *SDs*. Neither of these groups differed significantly from children who never attended center EEC, although it is important to reiterate that the no center care as well as the only early center care groups were very small, with limited statistical power to detect differences with the two larger groups of center preschool only and infant/toddler as well as preschool center care.

Alternative Model Specifications

A series of additional model specifications were estimated to test the robustness of the main effect results to a variety of concerns. First, models were rerun with the covariates entered separately at each wave and again with time-varying covariates averaged over Waves 1 through 3. Results were nearly identical to models with the covariates aggregated to highlight instability in children's home contexts. Second, models were specified without the propensity score weights, including the full set of covariates, child lags, and sample weights. Results were very similar to those presented in Tables 2 and 3, with a few instances in which the OLS results were slightly stronger than the PSW results, suggesting that the propensity score weights helped to adjust for selection bias.

Moderation by Family Socioeconomic Resources

Following the main effects models we assessed whether associations between EEC experiences and children's cognitive and behavioral skills differed for children from more versus less advantaged families. Moderation models were estimated for each of the three indicators of EEC experiences using first, a continuous measure of household income (averaged over Waves 1 through 3); second, categorical indicators of parent's highest level of educational attainment; and third, a continuous measure of home cognitive stimulation (averaged over Waves 2 and 3). We also assessed interactions with a dichotomous income measure delineating low-income children. Results (available upon request) did not show evidence suggesting that a greater duration, greater intensity, or different patterns of

center EEC were more or less beneficial for children's cognitive skills or behavioral functioning across family socioeconomic status. In each set of interactions, significant interactions occurred at or below the level expected by chance.

Moderation by Child Temperament

A final set of models considered whether the associations between EEC experiences and children's later functioning differed as a function of child temperament. These models were estimated using a continuous measure of child temperament at Wave 1. Results did not find that EEC experiences were linked to children's later cognitive and academic skills differently for children with easier versus more challenging temperaments (interaction results available upon request).

Discussion

As governments across many countries increasingly promote accessible, affordable, and high quality EEC programs in order to support parental employment and prepare children for school, greater attention is being drawn to the repercussions of EEC experiences for children's long-term cognitive and behavioral development. An extensive and robust body of research has assessed how attending EEC programs is predictive of children's core cognitive and behavioral functioning following school entry; however the vast majority of the findings derive from U.S. studies with little replication in other countries. As such, we have limited knowledge concerning other policy models of EEC and the generalizability of EEC effects across different populations and in diverse policy and cultural environments.

With a similar economic structure and use of both private and public provision of EEC as in the United States, Australia offers a context to replicate this research. In this article, we assessed links between center-based EEC and children's cognitive skills and behavioral functioning in first grade in a nationally representative sample of Australian children, attending to the duration, intensity, and developmental timing of children's experiences in center EEC programs. Incorporating numerous analytic techniques to adjust for selection bias, analyses found that greater duration and intensity of center EEC from infant through preschool years were linked with small enhancements in children's nonverbal fluid intelligence skills according to direct assessments but were not associated with children's vocabulary or academic (language, literacy, and math) skills. Greater duration and intensity of center EEC exposure was also predictive of small detriments to behavioral functioning, with children evincing lower attention skills, higher conduct problems, and lower prosocial behaviors according to both parent and teacher reports. Time in informal EEC arrangements as well as exposure to multiple concurrent EEC arrangements, in contrast, were not associated with children's later cognitive or behavioral functioning.

An examination of the timing of center care experienced by children provided some evidence suggesting that the results were being driven by children who were in center care during infancy/toddlerhood as well as preschool years; these children showed more advanced nonverbal fluid intelligence but also heightened conduct problems in first grade in comparison to their peers who had attended center programs only during preschool. Children who attended center-based EEC programs only during infancy/toddler-

hood did not differ significantly from their peers, although we caution that this group was extremely small, as the majority of Australian children attend preschool programs. The limited statistical power from small sample sizes made it difficult to properly model the effect of not attending center-based preschool.

Together, the combined results from the models assessing the duration, intensity, and timing of center EEC suggest that an accumulation of center care over both earlier and later years prior to school entry as well as higher hours of center care were driving both the enhancements in fluid intelligence as well as the detriments in behavioral and social skills found in children as they entered middle childhood. Prior to discussing the implications of these results, it is important to note limitations. Data limitations that were inherent in our reliance on panel data from the LSAC included incomplete information on all care settings attended and information on EEC only at distinct developmental periods rather than a continuous accounting of EEC use from birth through kindergarten entry. In addition, we were not able to assess the quality of children's EEC experiences. Although EEC quality is an essential concern of policy makers and practitioners, it is important to reiterate that a number of studies have found the type and duration of EEC children experience to be stronger predictors of children's functioning than EEC quality as assessed using current measures (Coley, Lombardi, et al., 2013; McCartney et al., 2010; NICHD ECCRN, 2003, 2006; Peisner-Feinberg et al., 2001; Votruba-Drzal et al., 2013). Moreover, scholars are questioning the validity of many existing measures of EEC quality (Gordon, Fujimoto, Kaestner, Korenman, & Abner, 2013; Sabol, Soliday Hong, Pianta, & Burchinal, 2013), highlighting the need for measurement development and a reexamination of components of EEC experiences most promotive of children's successful development. Finally, we reiterate that these data were correlational, precluding us from drawing truly causal inferences; yet the results emerged controlling for a wide range of child, parent, and family characteristics as well as earlier child cognitive/behavioral functioning, factors that may influence both children's differential selection into EEC as well as their later functioning. Moreover, our models incorporated propensity score weighting (PSW) techniques designed to better control for preexisting differences in children and families and hence identify less biased connections between EEC experiences and children's later functioning (Imbens, 2000).

It is also important to consider the size and practical significance of the findings. The effect sizes of center-based EEC duration, intensity, and timing on children's cognitive skills and behaviors were consistently small, ranging from .06 to .20 *SDs*. These effect sizes are relatively comparable to those reported in much of the EEC research from the United States, which are most often in the .10 to .20 *SD* range (Coley, Lombardi, et al., 2013; Loeb et al., 2007), although many of the U.S. analyses assessed child functioning during kindergarten, a shorter lag time than used in the current research assessing children's functioning in first grade. One manner of assessing the importance of such effect sizes is to compare them to effects of other variables. In the current research, for example, effects of EEC were of a similar size to the difference between parents with a college degree versus a high school degree, which was associated with a .25 *SD* shift in children's matrix reasoning skills and shifts of .07 *SD* and .04 *SD* in parent reports of attention skills and teacher reports of conduct problems, respectively, in the multivariate models adjusting for EEC and other child

and family covariates. EEC effects were also similar in size to a 1.0 *SD* increment in the cognitive stimulation provided in children's home environments, which predicted increments of .04 to .06 *SDs* in children's cognitive and behavioral skills.

Another mechanism for considering the practical importance of findings is to address the probability of longer-term repercussions. Recent long-term follow-ups of EEC effects in the United States have reported evidence that negative effects of center-based EEC on behavior problems, although small, remained relatively stable through elementary school (Belsky et al., 2007). Other work has suggested that although measurable effects on achievement may fade out during middle childhood, long-term benefits reemerge in adulthood (Deming, 2009), suggesting that even small effects on cognitive and behavioral skills in early childhood may have long-term consequences for children. Scholars have hypothesized that such "sleeper" effects may be due to noncognitive skills such as task persistence or social skills (Chetty et al., 2011; Deming, 2009) that in turn affect long-term educational and economic outcomes.

Together, results from this study replicate and extend prior research from the United States arguing that extensive center-based EEC may provide both benefits (to fluid intelligence skills) and risks (to behavioral skills) for children's later development (e.g., Coley, Lombardi, et al., 2013; Magnuson et al., 2007; NICHD ECCRN, 2003; Phillips et al., 2006; Votruba-Drzal et al., 2013). Our results extend this literature by carefully attending to duration, intensity, and timing of center-based care, finding that it is an accumulation of center exposure over developmental periods and with greater intensity that matters most for both cognitive skills and behavioral functioning. Also replicating U.S. research, exposure to more informal and home-based care settings was not significantly associated with children's later functioning, nor was use of multiple concurrent EEC arrangements.

Hence, one central message from this research concerns the replication of the general pattern of associations between center EEC and children's functioning, replication that is notable given the differences in EEC policy and accessibility. Though Australia has greater regulations of EEC quality standards and offers more direct and subsidized funding for EEC, center child care appears to hold generally similar repercussions for both Australian and American children. Together, results from this work echo recent calls for the need to further delineate mechanisms and to develop and replicate center-based EEC models that best support children's cognitive skills while also promoting successful behavioral development. Research assessing publicly supported American EEC programs, including public pre-K programs (Gormley et al., 2011) as well as Head Start (U.S. Department of Health and Human Services, Administration for Children and Families, 2010), has found that socioemotional functioning is not compromised when children attend high quality school-based pre-K programs. A variety of potential mechanisms have been hypothesized to explain these findings, including high overall program quality, high levels of teacher education and teacher salaries, as well as extensive attention to classroom management (Gormley et al., 2011; see also Raver et al., 2009). The Australian government recently committed to making access to part-time center-based preschool with qualified teachers universal and to implement a nationally regulated quality assessment system (Council of Australian Governments, 2014). These changes may offer an opportunity to further study

and identify promising EEC practices to best support children's healthy development.

Results from this study also provide some points of divergence from studies with American children. Most notably, links between EEC and children's cognitive skills were narrower than has been shown in research with American samples: Here, results emerged only for children's nonverbal fluid intelligence and not their receptive vocabulary or general academic skills (which included language, literacy, and math skills). Much of the prior U.S. research has considered direct assessments of children's math and reading skills rather than relying on teacher reports as the LSAC did, which may be less valid and reliable (e.g., Duncan & NICHD ECCRN, 2003; Gormley et al., 2005; Loeb et al., 2007; Magnuson et al., 2004; Votruba-Drzal et al., 2013).

The nonsignificant associations between EEC and children's language, literacy, and math skills in this sample also may be related to a second arena of divergence with results from American samples: the lack of differential associations between EEC and children's skills as a function of family socioeconomic status. One possible explanation for the small or neutral effects of EEC on children's cognitive and behavioral skills is that these small average effects are hiding significant heterogeneity. For example, prior work from the United States has found that center EEC is significantly beneficial for cognitive skills development in children of parents with lower incomes, less education, and lower provision of cognitive stimulation in their home environments while showing neutral or even slightly negative effects on the cognitive skills of children from more advantaged families (Loeb et al., 2004, 2007; McCartney et al., 2007; Votruba-Drzal et al., 2013; but see Belsky et al., 2007 and other NICHD work that has not found income moderation). In the LSAC sample, in contrast, we found no evidence for moderation of EEC effects by family income, parent education, or home cognitive stimulation, suggesting that the small benefits for children's fluid intelligence and risks for children's behavioral skills derived from greater exposure to center EEC were shared broadly across Australian children.

What might explain these different patterns of results? Recent cross-national work has argued that income inequality and poverty are notably lower in Australia than in the United States and further that differentials in child functioning related to family income, particular differentials in cognitive skills, are also lower in Australia (Bradbury, Corak, Waldfogel, & Washbrook, 2012). If there are narrower gaps in children's socioeconomic resources as well as narrower gaps in children's cognitive skills related to socioeconomic resources in Australia, then EEC in Australia will have a much weaker potential to help close such gaps. This line of reasoning helps to explain both the limited total effects of EEC on children's language and academic skills in the LSAC sample, as well as the lack of EEC by SES moderation. We reiterate the lack of moderation by children's temperament as well, furthering the argument that the small effects unearthed in this study were shared broadly across subgroups of children.

In conclusion, our analysis sought to replicate and extend prior EEC research by incorporating sophisticated modeling techniques to help adjust for selection bias in delineating prospective associations between EEC and children's functioning among children in Australia, a country with higher EEC attendance and quality standards, and a more homogenous population than the United States. Our results largely replicated prior results with American samples

of children suggesting that greater duration and intensity of center EEC from infancy through preschool is linked with small advantages in nonverbal fluid intelligence skills and small detriments in behavioral skills for Australian children. These effects were not found to be differentiated by family socioeconomic characteristics or child temperament suggesting that EEC has small average effects for Australian children. As Australia and the United States both seek to expand access to high-quality EEC programs, results from this work add to the growing literature suggesting the need to develop and replicate center-based EEC models that best support children's cognitive skills while also promoting successful behavioral development.

References

- Australian Bureau of Statistics. (2006). *Child care Australia* (Cat. no. 4402.0). Canberra, Australia: Author.
- Australian Government Department of Human Services. (2014, July 10). *Parenting payment*. Retrieved from <http://www.humanservices.gov.au/customer/services/centrelink/parenting-payment>
- Australian Government Family Assistance Office. (2011). *Child care benefit*. Retrieved from <http://www.familyassist.gov.au/payments/family-assistance-payments/child-care-benefit/>
- Belsky, J., & Pluess, M. (2012). Differential susceptibility to long-term effects of quality of child care on externalizing behavior in adolescence? *International Journal of Behavioral Development*, 36, 2–10. doi:10.1177/0165025411406855
- Belsky, J., Vandell, D. L., Burchinal, M., Clarke-Stewart, K. A., McCartney, K., Owen, M. T., & NICHD Early Child Care Research Network. (2007). Are there long-term effects of early child care? *Child Development*, 78, 681–701. doi:10.1111/j.1467-8624.2007.01021.x
- Bradbury, B., Corak, M., Waldfogel, J., & Washbrook, E. (2012). Inequality in early childhood outcomes. In J. Ermisch, M. Jantti, & T. Smeeding (Eds.), *From parents to children* (pp. 87–119). New York, NY: Russell Sage Foundation.
- Bradley, R. H., McKelvey, L. M., & Whiteside-Mansell, L. (2011). Does the quality of stimulation and support in the home environment moderate the effect of early education programs? *Child Development*, 82, 2110–2122. doi:10.1111/j.1467-8624.2011.01659.x
- Cabell, S. Q., Justice, L. M., Zucker, T. A., & Kilday, C. R. (2009). Validity of teacher report for assessing the emergent literacy skills of at-risk preschoolers. *Language, Speech, and Hearing Services in Schools*, 40, 161–173. doi:10.1044/0161-1461(2009/07-0099)
- Cain, G. (1975). Regression and selection models to improve nonexperimental comparisons. In C. A. Bennett & A. A. Lumsdaine (Eds.), *Evaluation and experiments: Some critical issues in assessing social programs* (pp. 297–317). New York, NY: Academic Press. doi:10.1016/B978-0-12-088850-4.50009-4
- Chetty, R., Friedman, J. N., Hilger, N., Saez, E., Schanzenbach, D. W., & Yagan, D. (2011). How does your kindergarten classroom affect your earnings? Evidence from Project Star. *The Quarterly Journal of Economics*, 126, 1593–1660. doi:10.1093/qje/qjr041
- Claessens, A., & Chen, J. (2013). Multiple child care arrangements and child well-being: Early care experiences in Australia. *Early Childhood Research Quarterly*, 28, 49–61. doi:10.1016/j.ecresq.2012.06.003
- Coley, R. L., Li-Grining, C., & Chase-Lansdale, P. L. (2006). Low-income families' child care experiences: Meeting the needs of children and families. In N. Cabrera, R. Hutchins, & E. Peters (Eds.), *From welfare to child care: What happens to children when mothers exchange welfare for work* (pp. 149–170). Mahwah, NJ: Erlbaum.
- Coley, R. L., Lombardi, C. M., Sims, J., & Votruba-Drzal, E. (2013). Early education and care experiences and cognitive skills development: A comparative perspective between Australian and American children. *Family Matters*, 93, 36–49.

Coley, R. L., Votruba-Drzal, E., Collins, M. A., & Miller, P. (2014). Selection into early education and care settings: Differences by developmental status. *Early Childhood Research Quarterly*, 29, 319–332. doi:10.1016/j.ecresq.2014.03.006

Coley, R. L., Votruba-Drzal, E., Miller, P., & Koury, A. (2013). Timing, type, and extent of child care and children's behavioral functioning in kindergarten. *Developmental Psychology*, 49, 1859–1873. doi:10.1037/a0031251

Côté, S. M., Borge, A. I., Geoffroy, M., Rutter, M., & Tremblay, R. E. (2008). Nonmaternal care in infancy and emotional/behavioral difficulties at 4 years old: Moderation by family risk characteristics. *Developmental Psychology*, 44, 155–168. doi:10.1037/0012-1649.44.1.155

Council of Australian Governments. (2014, June 10). *Early childhood*. Retrieved from http://www.coag.gov.au/early_childhood

Deming, D. (2009). Early childhood intervention and life-cycle skill development: Evidence from Head Start. *American Economic Journal: Applied Economics*, 1, 111–124.

Detting, A. C., Gunnar, M. R., & Donzella, B. (1999). Cortisol levels of young children in full-day childcare centers: Relations with age and temperament. *Psychoneuroendocrinology*, 24, 519–536. doi:10.1016/S0306-4530(99)00009-8

Dowling, A., & O'Malley, K. (2009). *Preschool education in Australia* (Policy brief). Australian Council for Educational Research. Retrieved from [Research.acer.edu.au/policy_briefs/1](http://research.acer.edu.au/policy_briefs/1)

Dowsett, C. J., Huston, A. C., Imes, A. E., & Gennetian, L. (2008). Structural and process features in three types of child care for children from high and low income families. *Early Childhood Research Quarterly*, 23, 69–93. doi:10.1016/j.ecresq.2007.06.003

Duncan, G. J., Magnuson, K. A., & Ludwig, J. (2004). The endogeneity problem in developmental studies. *Research in Human Development*, 1, 59–80. doi:10.1080/15427609.2004.968330

Duncan, G. J., & NICHD Early Child Care Research Network. (2003). Modeling the impacts of child care quality on children's preschool cognitive development. *Child Development*, 74, 1454–1475. doi:10.1111/1467-8624.00617

Dunn, L. M., & Dunn, L. M. (1997). *Examiner's manual for the PPVT-III: Peabody Picture Vocabulary Test* (3rd ed.). Circle Pines, MN: American Guidance Service.

Early, D. M., & Burchinal, M. R. (2001). Early childhood care: Relations with family characteristics and preferred care characteristics. *Early Childhood Research Quarterly*, 16, 475–497. doi:10.1016/S0885-2006(01)00120-X

Entwistle, D. R., & Alexander, K. L. (1993). Entry into school: The beginning school transition and educational stratification in the United States. *Annual Review of Sociology*, 19, 401–423. doi:10.1146/annurev.so.19.080193.002153

Fullard, W., McDevitt, S. C., & Carey, W. B. (1984). Assessing temperament in one-to three-year-old children. *Journal of Pediatric Psychology*, 9, 205–217. doi:10.1093/jpepsy/9.2.205

Fuller, B., Kagan, S. L., Loeb, S., & Chang, Y. (2004). Child care quality: Centers and home settings that serve poor families. *Early Childhood Research Quarterly*, 19, 505–527. doi:10.1016/j.ecresq.2004.10.006

Geoffroy, M., Côté, S. M., Giguère, C. É., Dionne, G., Zelazo, P. D., Tremblay, R. E., . . . Séguin, J. R. (2010). Closing the gap in academic readiness and achievement: The role of early childcare. *Journal of Child Psychology and Psychiatry*, 51, 1359–1367. doi:10.1111/j.1469-7610.2010.02316.x

Goodman, R. (1997). The Strengths and Difficulties Questionnaire: A research note. *Child Psychology & Psychiatry & Allied Disciplines*, 38, 581–586. doi:10.1111/j.1469-7610.1997.tb01545.x

Gordon, R. A., Fujimoto, K., Kaestner, R., Korenman, S., & Abner, K. (2013). An assessment of the validity of the ECERS-R with implications for measures of child care quality and relations to child development. *Developmental Psychology*, 49, 146–160. doi:10.1037/a0027899

Gormley, T., & Gayer, T. (2005). Promoting school readiness in Oklahoma: An evaluation of Tulsa's pre-K program. *The Journal of Human Resources*, 40, 533–558.

Gormley, W. T., Gayer, T., Phillips, D., & Dawson, B. (2005). The effects of universal pre-K on cognitive development. *Developmental Psychology*, 41, 872–884. doi:10.1037/0012-1649.41.6.872

Gormley, W. T., Phillips, D. A., Newmark, K., Welti, K., & Adelstein, S. (2011). Social-emotional effects of early childhood education programs in Tulsa. *Child Development*, 82, 2095–2109. doi:10.1111/j.1467-8624.2011.01648.x

Gray, M., & Sanson, A. (2005). Growing up in Australia: The Longitudinal Study of Australian Children. *Family Matters*, 72, 4–9.

Harrison, L., & Ungerer, J. (2005). What can the Longitudinal Study of Australian Children tell us about infants' and 4 to 5-year-olds' experiences of early childhood education and care? *Family Matters*, 72, 26–35.

Harrison, L., Ungerer, J., Smith, G., Zubrick, S., Wise, S., Press, F., . . . the LSAC Research Consortium. (2009). *Child care in Australia: An analysis of the Longitudinal Study of Australian Children* (Social Policy Research Paper No. 40). Retrieved from www.fahcsia.gov.au/sites/default/files/documents/05_2012/sprp_40.pdf

Imbens, G. W. (2000). The role of the propensity score in estimating dose-response functions. *Biometrika*, 87, 706–710. doi:10.1093/biomet/87.3.706

Kilday, C. R., Kinzie, M. B., Mashburn, A. J., & Whittaker, J. V. (2012). Accuracy of teacher judgments of preschoolers' math skills. *Journal of Psychoeducational Assessment*, 30, 148–159. doi:10.1177/0734282911412722

Leon, A. C., & Hedeker, D. (2007). A comparison of mixed-effects quantile stratification propensity adjustment strategies for longitudinal treatment effectiveness analyses of continuous outcomes. *Statistics in Medicine*, 26, 2650–2665. doi:10.1002/sim.2732

Li-Grining, C. P., Votruba-Drzal, E., Maldonado-Carreño, C., & Haas, K. (2010). Children's early approaches to learning and academic trajectories through fifth grade. *Developmental Psychology*, 46, 1062–1077. doi:10.1037/a0020066

Loeb, S., Bridges, M., Bassok, D., Fuller, B., & Rumberger, R. W. (2007). How much is too much? The influence of preschool centers on children's social and cognitive development. *Economics of Education Review*, 26, 52–66. doi:10.1016/j.econedurev.2005.11.005

Loeb, S., Fuller, B., Kagan, S. L., & Carroll, B. (2004). Child care in poor communities: Early learning effects type, quality, and stability. *Child Development*, 75, 47–65. doi:10.1111/j.1467-8624.2004.00653.x

Maccoby, E. E., & Lewis, C. C. (2003). Less day care or different day care? *Child Development*, 74, 1069–1075. doi:10.1111/1467-8624.00592

Magnuson, K. A., Meyers, M. K., Ruhm, C., & Waldfogel, J. (2004). Inequality in preschool education and school readiness. *American Educational Research Journal*, 41, 115–157. doi:10.3102/00028312041001115

Magnuson, K., Ruhm, C., & Waldfogel, J. (2007). The persistence of preschool effects: Do subsequent classroom experiences matter? *Early Childhood Research Quarterly*, 22, 18–38. doi:10.1016/j.ecresq.2006.10.002

Magnuson, K. A., & Votruba-Drzal, E. (2009). Enduring influences of childhood poverty. In S. Danziger & M. Cancian (Eds.), *Changing poverty* (pp. 153–179). New York, NY: Russell Sage Foundation.

McCartney, K., Burchinal, M., Clarke-Stewart, A., Bub, K. L., Owen, M. T., Belsky, J., & NICHD Early Child Care Research Network. (2010). Testing a series of causal propositions relating time in child care to children's externalizing behavior. *Developmental Psychology*, 46, 1–17. doi:10.1037/a0017886

McCartney, K., Dearing, E., Taylor, B. A., & Bub, K. L. (2007). Quality child care supports the achievement of low-income children: Direct and indirect pathways through caregiving and the home environment. *Journal of Applied Developmental Psychology*, 28, 411–426. doi:10.1016/j.appdev.2007.06.010

Meyers, M. K., & Jordan, L. P. (2006). Choice and accommodation in parental child care decisions. *Community Development*, 37, 53–70. doi: 10.1080/15575330609490207

Michel, S. (2003). Roots and branches: Comparing child care policymaking in the US and Australia. *Australian Journal of Early Childhood*, 28, 1–6.

Morrissey, T. W. (2009). Multiple child-care arrangements and young children's behavioral outcomes. *Child Development*, 80, 59–76. doi:10.1111/j.1467-8624.2008.01246.x

Morrissey, T. W. (2010). Sequence of child care type and child development: What role does peer exposure play? *Early Childhood Research Quarterly*, 25, 33–50. doi:10.1016/j.ecresq.2009.08.005

National Center for Education Statistics. (2002). *Early Childhood Longitudinal Study—Kindergarten class of 1998–99 (ECLS-K): Psychometric report for kindergarten through first grade* (NCES 2002–05). Washington, DC: U.S. Department of Education.

NICHD Early Child Care Research Network. (2002). Early child care and children's development prior to school entry: Results from the NICHD Study of Early Child Care. *American Educational Research Journal*, 39, 133–164. doi:10.3102/00028312039001133

NICHD Early Child Care Research Network. (2003). Does amount of time spent in child care predict socioemotional adjustment during the transition to kindergarten? *Child Development*, 74, 976–1005. doi:10.1111/j.1467-8624.00582

NICHD Early Child Care Research Network. (2005). *Child care and child development: Results from the NICHD Study of Early Child Care and Youth Development*. New York, NY: Guilford Press.

NICHD Early Child Care Research Network. (2006). Child care effect sizes for the NICHD Study of Early Child Care and Youth Development. *American Psychologist*, 61, 99–116. doi:10.1037/0003-066X.61.2.99

Organization for Economic Cooperation and Development. (2006). *Starting strong II: Early childhood education and care*. Paris, France: Author.

Pearl, J. (2009). Causal inference in statistics: An overview. *Statistics Surveys*, 3, 96–146. doi:10.1214/09-SS057

Peisner-Feinberg, E. S., Burchinal, M., Clifford, R. M., Culkin, M., Howes, C., Kagan, S. L., & Yazejian, N. (2001). The relation of preschool child care quality to children's cognitive and social developmental trajectories through second grade. *Child Development*, 72, 1534–1553. doi:10.1111/j.1467-8624.00364

Phillips, D. A., McCartney, K., & Sussman, A. L. (2006). Child care and early development. In K. McCartney & D. Phillips (Eds.), *Blackwell handbook of early childhood development* (pp. 471–489). New York, NY: Blackwell. doi:10.1002/9780470757703.ch23

Pluess, M., & Belsky, J. (2010). Differential susceptibility to parenting and quality child care. *Developmental Psychology*, 46, 379–390. doi:10.1037/a0015203

Raver, C., Jones, S., Li-Lining, C., Zhai, F., Metzger, M., & Solomon, B. (2009). Targeting children's behavior problems in preschool classrooms: A cluster-randomized controlled trial. *Journal of Consulting and Clinical Psychology*, 77, 302–316. doi:10.1037/a0015302

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70, 41–55. doi:10.1093/biomet/70.1.41

Royston, P. (2004). Multiple imputation of missing values. *The Stata Journal*, 4, 227–241.

Royston, P. (2005). Multiple imputation of missing values: Update of ice. *The Stata Journal*, 5, 527–536.

Sabol, T. J., Soliday Hong, S. L., Pianta, R. C., & Burchinal, M. R. (2013). Can rating pre-K programs predict children's learning? *Science*, 341, 845–846. doi:10.1126/science.1233517

Soloff, C., Lawrence, D., & Johnstone, R. (2005). *Sample design* (LSAC Technical Paper No. 1). Melbourne, Australia: Australian Institute of Family Studies.

Soloff, C., Lawrence, D., Misson, S., & Johnstone, R. (2006). *Wave 1 weighting and non-response* (LSAC Technical Paper No. 3). Melbourne, Australia: Australian Institute of Family Studies.

Strickland, J., Hopkins, J., & Keenan, K. (2012). Mother–teacher agreement on preschoolers' symptoms of ODD and CD: Does context matter? *Journal of Abnormal Child Psychology*, 40, 933–943. doi:10.1007/s10802-012-9622-y

U.S. Department of Health and Human Services, Administration for Children and Families. (2010). *Head Start Impact Study: Final report*. Washington, DC: Author.

U.S. General Accounting Office. (1997). *Welfare reform: Implications of increased work participation for child care* (GAO/HEHS 97–75). Washington, DC: Author.

Vandell, D. L., Belsky, J., Burchinal, M., Steinberg, L., & Vandergift, N. (2010). Do effects of early child care extend to age 15 years? Results from the NICHD Study of Early Child Care and Youth Development. *Child Development*, 81, 737–756. doi:10.1111/j.1467-8624.2010.01431.x

Vermeer, H. J., & van IJzendoorn, M. H. (2006). Children's elevated cortisol levels at daycare: A review and meta-analysis. *Early Childhood Research Quarterly*, 21, 390–401. doi:10.1016/j.ecresq.2006.07.004

Votruba-Drzal, E., Coley, R. L., Koury, A., & Miller, P. (2013). Center-based child care and academic skills development: Importance of timing and household resources. *Journal of Educational Psychology*, 105, 821–838. doi:10.1037/a0032951

Waldfogel, J. (2009). The role of family policies in anti-poverty policy. *Focus*, 26(2), 50–55. Retrieved from <http://www.irp.wisc.edu/publications/focus/pdfs/foc262i.pdf>

Watamura, S. E., Donzella, B., Alwin, J., & Gunnar, M. R. (2003). Morning-to-afternoon increases in cortisol concentrations for infants and toddlers at child care: Age differences and behavioral correlates. *Child Development*, 74, 1006–1020. doi:10.1111/1467-8624.00583

Wechsler, D. (2004). *The Wechsler Intelligence Scale for Children—Fourth edition*. London, United Kingdom: Pearson Assessment.

Wetherby, A. M., & Prizant, B. M. (2001). *Communication and Symbolic Behavior Scales Developmental Profile: Infant-Toddler Checklist*. Baltimore, MD: Brookes.

Yamauchi, C., & Leigh, A. (2011). Which children benefit from non-parental care? *Economics of Education Review*, 30, 1468–1490. doi:10.1016/j.econedurev.2011.07.012

Zigler, E., Marsland, K., & Lord, H. (2009). *The tragedy of child care in America*. New Haven, CT: Yale University Press.

Appendix

Estimating Each Child's Propensity to Be in Center EEC

Variable	Duration of care	Intensity of care	Patterns of care		
	% waves center care	Avg. hours in center care	Early center, no preschool	Preschool only	Early center plus preschool
Wave 1 predictors					
Child age	0.004 (0.001)**	0.009 (0.004)*	0.116 (0.053)*	-0.061 (0.029)*	-0.007 (0.028)
Child male	-0.003 (0.008)	-0.020 (0.022)	0.010 (0.275)	0.189 (0.148)	0.109 (0.151)
Child low birth weight	-0.011 (0.017)	0.024 (0.047)	-0.722 (0.742)	0.019 (0.325)	-0.135 (0.315)
Child bad health	0.059 (0.027)*	0.164 (0.076)*	-0.338 (1.059)	-0.022 (0.495)	0.289 (0.497)
Child cognitive skills	-0.002 (0.000)**	-0.008 (0.001)**	0.000 (0.016)	0.001 (0.009)	-0.016 (0.009) [†]
Child temperament	0.000 (0.007)	0.022 (0.020)	0.156 (0.235)	0.085 (0.135)	0.076 (0.134)
Parent Asian	-0.059 (0.017)**	-0.094 (0.05) [†]	-0.237 (0.744)	-0.105 (0.325)	-0.401 (0.314)
Parent Aboriginal	-0.054 (0.021)*	0.010 (0.066)	0.148 (0.473)	-0.665 (0.285)*	-0.650 (0.268)*
Immigrant household	0.014 (0.010)	0.076 (0.031)*	-0.081 (0.378)	-0.096 (0.194)	0.014 (0.186)
Non-English household	-0.073 (0.015)**	-0.092 (0.046) [†]	-0.344 (0.598)	-0.286 (0.252)	-0.751 (0.249)**
Child number siblings	-0.039 (0.004)**	-0.095 (0.010)**	-0.177 (0.168)	-0.339 (0.065)**	-0.515 (0.062)**
Parent married	0.029 (0.010)**	-0.029 (0.030)	-0.344 (0.354)	0.480 (0.198)*	0.502 (0.192)*
Youngest parent's age	0.002 (0.001)*	0.005 (0.003) [†]	-0.030 (0.039)	0.025 (0.018)	0.013 (0.018)
Parent < high school education	-0.035 (0.027)	-0.118 (0.075)	-0.664 (0.860)	-0.327 (0.394)	-0.394 (0.418)
Parent some college	0.027 (0.021)	0.041 (0.059)	0.003 (0.826)	-0.206 (0.327)	0.114 (0.335)
Parent college/grad school	0.053 (0.021)*	0.128 (0.059)*	-0.022 (0.811)	0.029 (0.323)	0.403 (0.325)
Household income	0.000 (0.000)**	0.000 (0.000)**	0.000 (0.000)	0.000 (0.000)**	0.000 (0.000)**
Mother employment hours	0.004 (0.000)**	0.020 (0.001)**	0.004 (0.014)	-0.002 (0.008)	0.016 (0.008)*
<i>F</i> of model	28.75	34.35	5.72	5.72	5.72
<i>R</i> ² /Pseudo <i>R</i> ²	0.12	0.17			
Range of propensity scores	0.073–0.861	0.055–2.660	0.001–0.289	0.164–0.746	0.059–0.812

Note. EEC = early education and care; Avg. = average.

[†] *p* < .10. * *p* < .05. ** *p* < .01.

Received April 24, 2013
 Revision received June 11, 2014
 Accepted June 17, 2014 ■

“He Who Can, Does; He Who Cannot, Teaches?”: Stereotype Threat and Preservice Teachers

Toni A. Ihme
FernUniversität in Hagen

Jens Möller
Christian-Albrechts-Universität zu Kiel

Stereotype threat is defined as a situational threat that diminishes performance, originating from a negative stereotype about one's own social group. In 3 studies, we seek to determine whether there are indeed negative stereotypes of students who have chosen a career in teaching, and whether the performance of these students is affected by stereotype threat. Responses to open-ended questions (Study 1, $N = 82$) and comparisons in closed-ended response format (Study 2, $N = 120$) showed that preservice teachers are perceived as having a low level of competence and a high level of warmth, in keeping with the paternalistic stereotype. We conclude that a stereotype does indeed exist that attributes lower competence to prospective teachers. In Study 3 ($N = 262$), a group of preservice teachers was subjected to stereotype threat. In keeping with the stereotype threat model, that group performed worse on a cognitive test than the group of similar students who were not under stereotype threat; the performance of students in the field psychology did not differ in response to the threat condition. This study is the 1st to show the effects of stereotype threat on students preparing for a teaching career.

Keywords: preservice teachers, stereotype threat, stereotype, performance

The term *stereotype threat* refers to a situational threat that diminishes performance, originating from a negative stereotype about one's own social group (Steele, 1997; Steele & Aronson, 1995). Situational pressure results when members of a group find themselves in a situation that is associated with a negative stereotype of that group, and they are anxious about confirming the stereotype or being judged by it. This leads to an emotional response that impairs the individual's cognitive functioning and performance (Schmader, Johns, & Forbes, 2008). Considerable evidence has shown that female students perform less well than their male counterparts in mathematics and related fields. Girls who are reminded of their gender before taking a math test worry about confirming the stereotype that women are less capable than men in math and science. These negative feelings and thoughts consume some of the cognitive resources necessary for performing well on the test, leading to results that are worse than they would be otherwise (Huguet & Régner, 2007; Ihme & Mauch, 2007; Jordan & Lovett, 2007; Keller, 2007; Keller & Dauenheimer, 2003; Muzzatti & Agnoli, 2007, among others). In addition to lowering school achievement, stereotype threat interferes with learning itself (Rydell, Rydell, & Boucher, 2010; Rydell, Shiffrin, Boucher, Van Loo, & Rydell, 2010) and weakens identification

with the affected domain or group (e.g., Woodcock, Hernandez, Estrada, & Schultz, 2012).

Ethnic stereotypes, too, have been the subject of considerable study. McKown and Weinstein (2003) found that the concentration and working memory of African American and Latino students declined after they were reminded of the stereotype of these groups as intellectually inferior. Désert, Préaux, and Jund (2009) have shown how stereotype threat affects the performance of children from socially disadvantaged families on intelligence tests. Stereotype threat is an issue not only for students but also for individuals in the labor force. von Hippel, Kalokerinos, and Henry (2013), for example, have found that older workers who face age-related stereotypes are less satisfied with their jobs and more likely to quit.

Because negative stereotypes are associated with certain types of jobs, stereotype threat may play a role in that context as well. However, unlike other social groups (such as those related to gender or ethnicity), jobs are usually chosen. As a result, it is possible to leave profession-related groups. On the one hand, people might therefore regard their membership in these groups as less binding and less significant, and this might make them less vulnerable or even immune to stereotype threat. On the other hand, people might regard their membership as binding and significant precisely because of their choice (Fisher & Andrews, 1976), and this might make them vulnerable to stereotype threat.

The focus of our studies is to determine whether stereotype threat affects students training for a career in teaching. This is an important question, because stereotype threat may have a detrimental effect on learning and performance (Rydell et al., 2010; Taylor & Walton, 2011) and ultimately lead to disidentification with the field of study (and thus also the teaching profession), possibly even causing students to drop out (see Milner & Woolfolk Hoy, 2003, for a discussion of the possible effects of stereotype threat on African American teachers).

This article was published Online First July 7, 2014.

Toni A. Ihme, Department of Social Psychology, FernUniversität in Hagen; Jens Möller, Department of Educational Psychology, Christian-Albrechts-Universität zu Kiel.

Correspondence concerning this article should be addressed to Toni A. Ihme, Department of Social Psychology, FernUniversität in Hagen, Universitätstraße 33, 58097 Hagen, Germany. E-mail: Toni-Alexander.Ihme@fernuni-hagen.de

Stereotype Threat in the Teaching Profession and Teacher Training

Teachers and preservice teachers are subject to considerable negative stereotyping (Blömeke, 2005; Spinath, van Ophuysen, & Heise, 2005; Swetnam, 1992). George Bernard Shaw's comment that "he who can, does; he who cannot, teaches" reflects an attitude toward the teaching profession that is still common today (teaching, 2013). It is a view that is often found in the media as well (Blömeke, 2005; Swetnam, 1992). Even preservice teachers themselves seem to share these negative perceptions; a study by Carlsson and Björklund (2010) has shown that they, too, view preschool teachers as considerably less competent than lawyers (though high in warmth).

It therefore appears that teachers, more than other occupational groups, are viewed as less competent, and they are confronted with these negative stereotypes even during their training. It is widely believed that preservice teachers are weaker than other students in the qualities needed to earn an academic degree (such as intelligence and achievement motivation; Spinath et al., 2005). In their comparative study Spinath et al. (2005) looked at various fields of study (education science, economics, mathematics, natural sciences, and engineering, as well as teacher training programs geared to a variety of school types), finding no difference between the preservice teachers and other students in terms of intelligence, achievement motivation, or reading skills. However, Spinath and colleagues (2005) also concluded that despite the lack of real differences in cognitive ability, the very existence of these stereotypes can have a negative effect. Yet, no systematic studies have investigated the incidence and nature of stereotypes associated with preservice teachers or looked at the effects of stereotype threat on this population.

Our studies are intended to show whether preservice teachers believe that others view them as less competent and whether that stereotype leaves them vulnerable to stereotype threat. Our first step was to conduct two exploratory studies (Studies 1 and 2) to determine the general salience of these stereotypes. On the basis of the stereotype content model (Fiske, Cuddy, Glick, & Xu, 2002; Fiske, Xu, Cuddy, & Glick, 1999), we examined whether preservice teachers (Study 1) and other persons (Study 2) believe in the existence of the oft-cited stereotype that preservice teachers are less capable. Often research on stereotype threat simply assumes the existence of the investigated stereotype. This may be justified in certain cases (such as gender stereotypes on performance in mathematics, for example) but needs confirmation in others. The more a stereotype is widely held and consensual, the more likely it is to be chronically accessible: Widely held stereotypes may require only subtle hints to become salient (such as TV spots not even directly mentioning or showing the stereotype in question—e.g., in case of gender stereotypes). That is why we conducted two studies using different methods and inviting different participants.

Of central importance is our experimental study (Study 3) in which preservice teachers were confronted with this stereotype when taking a cognitive test. Their results were compared with those of other preservice teachers who were not subjected to stereotype threat, as well as with a control group.

Study 1

In the first study, consisting of open-ended questions, preservice teachers were asked to describe the stereotypes of their ingroup. Their responses were categorized according to the stereotype content model (Fiske et al., 1999, 2002), which classifies social stereotypes in terms of two dimensions—competence and warmth—using a four-quadrant matrix. Various combinations of these two qualities result in four stereotypes: The paternalistic stereotype is usually associated with individuals of lower status (housewives, older people) who do not represent competition for social resources; these individuals are perceived to have a low level of competence and a high level of warmth. "Admiration" attributes high scores on both dimensions to the respective groups (e.g., the ingroup). Groups that fit the contemptuous stereotype (e.g., the homeless) are viewed as incompetent and cold, whereas the envious stereotype applies to groups regarded as competent, but cold (e.g., the rich).

Given what we know about the negative stereotyping of preservice teachers and the stereotypes found in the literature, we would expect these students to fit the paternalistic stereotype most closely—in other words, to be viewed as low on competence but high on warmth.

Method

Sample. Members of the sample—students at a university in northern Germany ($N = 82$) who were enrolled in a master's program in teaching (academic track *Gymnasium*, $M = 9$ semesters, $SD = 1.78$)—in Study 1 were asked about stereotypes of their group. The average age of the respondents was $M = 26$ ($SD = 4.58$); 76.8% of them were women. Respondents were recruited through university courses.

To clarify the background of the preservice teacher samples in Study 1 and Study 3 (the sample in Study 2 consisted of students from other fields of study as well as participants from different nonuniversity backgrounds), we summarize a typical teacher training program in Germany. In most German federal states, teacher training at university is composed of a 3-year bachelor's (B.A./B.Sc.) program and a 2-year master's (M.Ed.) program, including the academic study of two scientific disciplines and didactics for the corresponding school subjects. Additionally, students take a variety of courses in educational sciences. During their time at the university, they attend the same courses as students not training to be teachers (e.g., preservice mathematics teachers attend together with students who want to become mathematicians). After completing the master's program, they start their on-the-job training in schools (*Referendariat*). After 1–2 years of training, they become proper teachers.

Procedure. The survey was conducted at the beginning of these courses. The students were given 5 min to answer the following question in a few key words: "In your opinion, what characteristics do other students ascribe to the 'typical student' preparing for a career in teaching?"

Other studies have used similar open-ended questions to generate stereotypes (e.g., Devine, 1989). In this case, the purpose was to determine how preservice teachers believe they are perceived by other students. The sample was limited to students preparing for a teaching career, because stereotype threat will only have an effect if participants are aware of a negative stereotype and assume that

they are being judged by that stereotype (Steele, 1997). We chose an open-ended format so that respondents could freely generate stereotypes.

Analysis. Study 1 recorded $N = 398$ statements (average $M = 4.84$ statements per person). Two student reviewers categorized responses on the basis of the two dimensions of the stereotype content model (competence, warmth). They used two five-step scales to rate each response in terms of competence (ranging from $-2 =$ *indicative of low competence* to $0 =$ *neutral* to $+2 =$ *indicative of high competence*) and warmth (ranging from $-2 =$ *indicative of low warmth* to $0 =$ *neutral* to $+2 =$ *indicative of high warmth*). Responses such as "incapable" or "not suited to academic studies" were consistently classified as indicating incompetence. Responses like "vain" or "politically on the left" were considered to be neutral. Descriptions such as "capable" and "striving" were viewed as indicative of competence.

To confirm that the two dimensions of competence and warmth formed the basis of reviewers' categorizations, we followed a three-step procedure. First, we randomly split the sample of statements. Second, a factor analysis of the four categorizations (one supposed competence and one supposed warmth rating per reviewer) made by the reviewers was conducted on one half the sample. The analysis resulted in two clearly distinguishable factors (all factor loadings > 1.73) explaining 87.20% of the variance. Both competence ratings formed one factor, whereas both warmth ratings formed the other one. Third, a confirmative factor analysis of both reviewers' categorizations on the competence and the warmth dimension (using the other half of the sample) produced an acceptable model fit, $\chi^2(1) = 1.78$, $p < .25$, root-mean-square error of approximation = .06, comparative fit index = 1.00, thus confirming the two-dimensional structure.

Results

We analyzed the mean levels of competence and warmth on the basis of the reviewers' ratings. If the two student reviewers disagreed, a third reviewer was asked to render a final decision that was included in the mean score. Because we applied two *t* tests to investigate the mean levels of competence and warmth attributed to preservice teachers, the Bonferroni method used to correct the alpha level resulted in an alpha level of .025. The mean level of competence attributed to these students was $M = -.38$ ($SD = .96$), a value significantly below the midpoint of the scale, $t(397) = -7.86$, $p < .001$, $d = .40$. The mean level of warmth was $M = .29$ ($SD = .97$). This value, too, deviated—in this case in a positive direction—from the midpoint of the scale, $t(397) = 6.07$, $p < .001$, $d = .30$. Thus, preservice teachers were perceived to have a low level of competence and a high level of warmth.

Discussion

The first study consisted of an open-ended survey aimed at identifying the stereotypes preservice teachers believe others attach to them. The stereotype content model was used to analyze their responses. Results showed that these students are aware that others view them as less competent, a negative stereotype that was identified by Spinath and colleagues (2005). Overall, the first study painted a picture of a less competent but sociable group, in keeping with the paternalistic stereotype.

However, because Study 1 was a qualitative survey of a very selective sample of members of the potentially stereotyped group, it is not able to shed light on the salience of the stereotype, nor does this kind of survey reveal just how negative these stereotypes are, relative to other groups. We therefore conducted another study with a larger sample, which allowed us to compare the group of preservice teachers with other groups of students in an effort to find additional empirical evidence of a stereotype that preservice teachers are less competent. This study was intended to determine the general salience of this negative stereotype and its relative level of negativity.

Study 2

The second study was conducted online, and various groups (students, employed individuals, etc.) were asked to describe the characteristics of students in several fields. The survey was designed in accordance with the stereotype content model (Fiske et al., 1999, 2002). We used a method established by Fiske et al. (1999) to assess these groups' competence and warmth (see below).

Method

Sample. The analysis included data on $N = 120$ individuals. The average age of the respondents was $M = 29$ ($SD = 9.92$; range = 16–62 years); 73.3% of them were women. The majority were university students ($n = 84$) of various fields of study. The remaining participants were either members of the workforce ($n = 22$), people without employment ($n = 11$), or pupils ($n = 3$). In keeping with Reips' (2002) recommendations for online research, respondents were included in the analysis only if it was clear that they had completed the survey without interruption. Thus, $n = 3$ individuals had been excluded prior to the analysis.

Procedure. To test the assumption that preservice teachers fit the paternalistic stereotype, we looked at students in four disciplines (teacher training, law, computer science, and psychology) in light of the dimensions of the stereotype content model, analogous to the work of Fiske et al. (1999). The respondents were given a list of the 27 characteristics used by Fiske and colleagues (1999), translated into German, for the purpose of identifying levels of competence and warmth.

Respondents were instructed as follows:

The purpose of this survey is to determine how students in four different fields (teacher training, law, computer science, and psychology) are perceived. We are not interested in your personal beliefs, but in how you think these groups are viewed by others.

This wording was modeled after Fiske et al. (1999) and intended to prevent respondents from answering in a way they considered socially acceptable. Respondents indicated the degree to which each characteristic applied to the respective group, using a 5-point Likert scale ranging from 0 (*not at all*) to 4 (*extremely*); the characteristics were listed in random order. On the basis of these responses, we estimated the degree to which each characteristic applied to students in the four groups.

Analysis. The first steps of our analysis follow Fiske et al. (1999). A principal components analysis of the 27 characteristics was conducted for each group (teaching, law, computer science,

and psychology students), for a total of four analyses. Out of the resulting factors, oblique rotation, we identified the ones on which the items "competent" (translated as "kompetent") and "likable" (translated as "sympathisch") loaded the highest (factor loadings of over .40). Then the characteristics were identified that consistently, for all four groups, loaded on the same factor (once again, factor loadings of over .40), either with "competent" or with "likable." In contrast to Fiske et al. (1999), we included only the characteristics that were present in the factor solutions for all of the groups. The item "competent" loaded with "industrious," "intelligent," and "determined." The item "likable" loaded with "helpful," "sincere," "warm," and "kind" (see Table 1). The resulting scales were sufficiently reliable ($.80 \leq \alpha \leq .85$).

We subsequently conducted within-subject analyses of variance (ANOVAs), comparing scale values for competence and warmth (see Table 2) for the four groups. Our central hypothesis suggests that preservice teachers are regarded as less competent than other students.

Results

The results show that competence ratings differ between different fields of study, $F(3, 357) = 52.55, p < .001, d = 1.33$. Paired comparisons (applying the Bonferroni method) were made to examine which pairs of means differed. Preservice teachers were perceived to be significantly less competent than the other groups (all $p < .001$). The other groups (psychology, law, and computer science students) showed no differences in perceived competence (all $p > .07$).

The results show that warmth ratings differ between different fields of study, $F(2.8, 336.4) = 132.78, p < .001, d = 2.10$.¹ Again, paired comparisons were made to examine which pairs of means differed. Preservice teachers were seen as scoring significantly higher on warmth than law or computer science students ($p < .001$), but this was not the case when they were compared with psychology students ($p = .65$). Psychology students, too, were seen as warmer than law or computer science students ($p < .001$). Furthermore, law students were considered to be significantly less warm than computer science students ($p < .001$).

A final t test for dependent samples showed that preservice teachers were regarded as significantly less competent than warm, $t(119) = -7.41, p < .001, d = .72$. From the perspective of the stereotype content model, they therefore fit the paternalistic stereotype, with low competence and high likability.

Table 1
Traits for Study 2

Competent	Warm	Arrogant	Determined
Likable	Gullible	Industrious	Tolerant
Helpful	Confident	Gentle	Complaining
Spineless	Dictatorial	Intelligent	Irritable
Sincere	Competitive	Good-natured	Egoistical
Cold	Independent	Kind	Passive
Hostile	Whiny	Greedy	

Note. These traits were adapted from Fiske, Cuddy, and Glick (1999) for the purpose of the present research and translated into German.

Table 2

Study 2: Means (and Standard Deviations) for Competence and Warmth of the Assessed Fields of Study

Dimension	Field of study	<i>M</i> (<i>SD</i>)
Competence	Teacher training	2.02 (.72)
	Law	2.90 (.71)
	Computer science	2.75 (.61)
	Psychology	2.73 (.74)
Warmth	Teacher training	2.60 (.58)
	Law	1.39 (.60)
	Computer science	1.84 (.57)
	Psychology	2.50 (.65)

Discussion

The open-ended questions posed in Study 1 revealed that perceptions of preservice teachers corresponded to the paternalistic stereotype. In Study 2, we compared those stereotypes with stereotypes of students in other fields. As expected, perceptions of preservice teachers conformed to the paternalistic stereotype proposed by the stereotype content model, with low scores for competence and higher ones for warmth. This is in keeping with the findings of Carlsson and Björklund (2010) in their study of preschool teachers. Our findings also confirmed the existence of a negative stereotype related to the competence of preservice teachers, as the literature suggests. The next question was whether the performance of these students is adversely affected when that stereotype is made salient. Study 3 was devoted to answering that question.

Study 3

In Study 3, we tested the hypothesis that competence-related stereotype threat leads to weaker performance by preservice teachers: Preservice teachers who are subjected to stereotype threat perform less well on a test of cognitive ability than preservice teachers who are not.

Method

Sample. Test participants included $N = 262$ preservice teachers (academic track *Gymnasium*, $n = 134$) and psychology students ($n = 128$) at a university in northern Germany ($M = 23.69$ years of age, $SD = 4.35$; 72.9% female) who were recruited through university courses. The psychology students were included as a control group.

Independent variables. The independent variables were the test condition (IV1: Stereotype threat present vs. not present) and the academic field of the test subjects (IV2: teaching vs. psychology). Participants were randomly assigned to the respective IV1 condition. Stereotype threat was conveyed through the test instructions. We consciously chose a moderately explicit approach (Nguyen & Ryan, 2008) by activating different fields of study as social category with no explicit mention of expectation of weaker performance for three reasons: First, in

¹ Because the Mauchly test showed a significant result, the Greenhouse-Geisser estimator was applied.

everyday life, stereotype threat is generally activated almost unnoticed (Steele, 2010). Second, as numerous studies have shown, activating stereotype threat does not require a direct link to the relevant stereotype (Davies, Spencer, Quinn, & Gerhardstein, 2002; Ihme & Mauch, 2007); however, the stereotype itself must be sufficiently salient. Third, we wanted to rule out the possibility of stereotype reactance, which occurs when individuals feel challenged by the explicit activation of a negative stereotype and respond by acting in a way that is contrary to the stereotype (Kray, Thompson, & Galinsky, 2001). A more subtle approach to activating the stereotype is intended to prevent that effect.

Psychology was selected as the comparison discipline based on the finding of Study 2 that psychology students and preservice teachers differ only in the stereotype content model's competence dimension and not in warmth. Psychology students, although similarly stereotyped in the warmth dimension and similar to preservice teachers in other characteristics (e.g., gender distribution), constitute a different group who is not targeted by the performance-related stereotype. When group membership is made salient, stereotype threat should affect the performance of preservice teachers only. In stereotype threat research, the choice of the comparison group is most often a trivial point. In case the threatened group is an ethnic minority, the comparison group is often the ethnic majority; in case the threatened group consists of female participants, the comparison group consists of male participants, and so on. However, in case of different fields of study, the answer is less straightforward. In the end, we decided for a rather similar comparison group so that any effect we found could be attributed to this one particular stereotype that we targeted.

Dependent variables. To test performance, we used the matrix subtest from the Intelligence Structure Test 2000R (Liepmann, Beauducel, Brocke, & Amthauer, 2007). It was chosen because it does not include mathematics or verbal items, which might activate subject-specific, school-related aspects of the study participants' self-concept. For each of the 20 items on the test, the participants were asked to select from five symbols the one that completes a pattern consisting of between four and nine symbols. They were given 10 min to complete this task, and could score a maximum of 20 points. The internal consistency of the test items in the sample was $\alpha = .64$.

Procedure. The test was conducted in groups of up to 50 participants in an introductory psychology course. All experimental sessions took place in the same lecture hall and at the same time of day. All participants in one session were either preservice teachers or students of psychology. After welcoming the participants, the investigator explained that they were to take a multiple-choice test and that further details about the purpose of the study could be found in the standardized instructions. Each participant was then given an envelope containing the test materials. The materials were distributed in envelopes so that the investigator and his assistants could not see which participant was assigned to which test condition. In each session, half of the materials were prepared for either of the experimental conditions (stereotype threat or not). The instructions found in the envelope contained the experimental manipulation. The manipulation for the stereotype threat condition stated:

Dear participant,

On the following pages you will find a test of cognitive ability. This questionnaire is intended to test recent research findings showing a significant difference between pre-service teachers and students in other fields (such as psychology, educational science² etc.) in their cognitive capabilities. We will be gathering a variety of cognitive data from you. Please concentrate carefully as you follow the instructions and complete these items to the best of your ability.

The instruction for the no-stereotype threat condition merely stated that the questionnaire was intended to "test recent research findings" without giving any further details or mentioning any differences in the cognitive abilities of students from different fields of study:

Dear participant,

On the following pages you will find a test of cognitive ability. This questionnaire is intended to test recent research findings. We will be gathering a variety of cognitive data from you. Please concentrate carefully as you follow the instructions and complete these items to the best of your ability.

Students in both disciplines (preservice teachers and psychology) within the experimental conditions received exactly the same instructions. After reading these general instructions, the participants read the specific test instructions and then completed the test. In a final step, they were asked to provide demographic information and to assess their consent to the manipulation check item. The experiment concluded with an extensive debriefing of the participants.

Manipulation check. To assess the effectiveness of the experimental manipulation, we asked participants about their beliefs concerning the performance differences between students from different fields of study: "Students of different fields of study will differ in their performance in this test." Consent to this item was measured on a 6-point scale anchored at the endpoints by the phrases *strongly disagree* (1) and *strongly agree* (6). This method resembles the manipulation checks in other studies in which the participants were asked about the purpose of the experiment, details of the instructions, or their belief in the activated stereotype (e.g., Alter, Aronson, Darley, Rodriguez, & Ruble, 2010).

Analysis. The manipulation check and the matrix test results were evaluated using a two-factor ANOVA (test condition: Non-stereotype threat condition = 0, stereotype threat condition = 1; field of study: preservice teachers = 0, students of psychology = 1).³ Simple effect tests were used to analyze the results of the matrix test in greater detail.

² The original German term applied in this instruction is *Pädagogik*. The English translation of the term implies a slightly different meaning than the original German term. *Lehramt* (the field of study for the preservice teachers in Germany) and *Pädagogik* are not the same. Whereas the first includes elements of the second, students of *Pädagogik* (in Germany) do not become teachers but social workers or workers in other areas of welfare.

³ Considering that a majority of the sample consisted of female students who might also be subject to the low-competence and high-warmth stereotype, we also tested for possible effects of gender. However, because we found none, these analyses are not reported here.

Results

The ANOVA of the manipulation check showed a significant main effect for the stereotype threat factor, $F(1, 252) = 4.22, p < .05, d = .29$. Participants subject to the stereotype threat condition ($M = 4.26, SD = 1.19$) were more likely than members of the control group to believe that students in different fields would perform differently on the matrix test ($M = 3.93, SD = 1.39$). No significant effects were found for the field of study, $F(1, 252) = .01, ns, d < .10$, or for the interaction of the factors, $F(1, 252) = .03, ns, d < .10$. The manipulation can therefore be regarded as successful.

Table 3 shows the mean values and standard deviations of the participants' test results (see also Figure 1). The ANOVA (two factors: stereotype threat and field of study) revealed a significant interaction, $F(1, 258) = 6.72, p < .01, d = .35$, and a significant main effect for the field of study, $F(1, 258) = 10.10, p < .01, d = .41$. The mean value of the preservice teachers' test results was lower than that of the psychology students. The main effect of stereotype threat was not significant, $F(1, 258) = 2.17, ns, d = .20$.

In order to explain this interaction, post hoc simple effects analyses were carried out for each of the independent variables. Inspection by the least significant difference test revealed that preservice teachers who had been subjected to stereotype threat performed worse on the test relative to preservice teachers who had not been subjected to stereotype threat, $F(1, 258) = 8.44, p < .01, d = .50$. There was no significant difference between psychology students in the stereotype threat condition and psychology students in the no-stereotype threat condition, $F(1, 258) = .62, ns, d = .14$. We also compared preservice teachers and psychology students in the same test condition. Preservice teachers who had been subjected to stereotype threat performed worse on the test relative to psychology students who had been subjected to stereotype threat, $F(1, 258) = 17.00, p < .01, d = .71$. There was no significant difference between preservice teachers in the no-stereotype threat condition and psychology students in the no-stereotype threat condition, $F(1, 258) = .17, ns, d = .07$.

In summary, preservice teachers who had been subjected to stereotype threat proved to be the only participants whose performance was affected by the experimental condition. The main effect of the field of study as identified by the ANOVA can therefore be attributed to the interaction between test condition and field of study.

Discussion

In Study 3, we examined whether the stereotype of lower cognitive capacity has a negative effect on preservice teachers' intelligence test performance. The results confirm the typical findings of stereotype threat research: Members of a group associated with a negative stereotype (preservice teachers) performed less

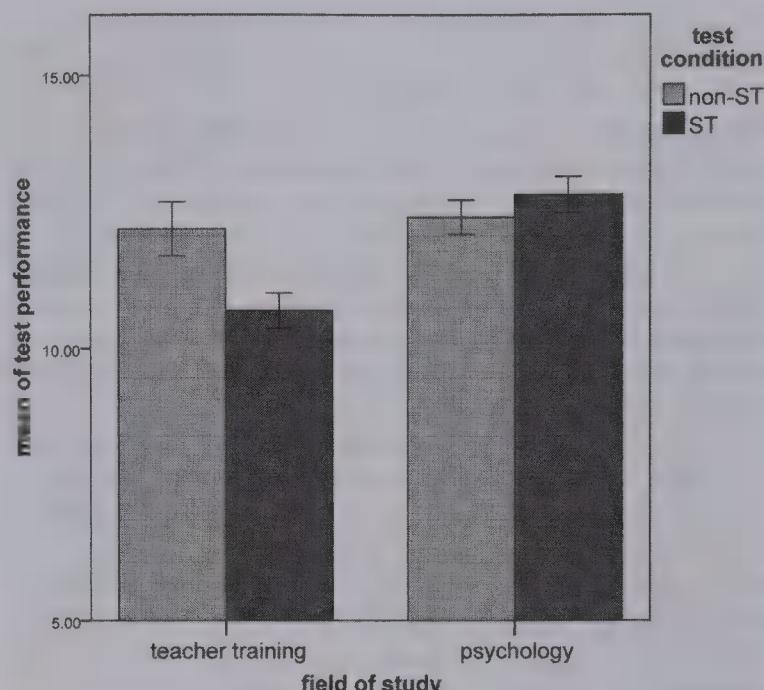


Figure 1. Means of test performance in Study 3 for each field of study and for each test condition (stereotype threat [ST]; no stereotype threat [non-ST]). Error bars represent standard errors.

well on a test of cognitive ability when they were subject to stereotype threat than members of the same group who were not. Stereotype threat had no effect on the performance of members of a group who is not associated with that stereotype (psychology students).

General Discussion

The studies described in this article expand on previous research on teacher education and the teaching profession. We have shown, first of all, that the stereotype commonly found in the literature does indeed exist, which views preservice teachers as less intellectually capable. Studies 1 and 2 demonstrated that the patriarchal stereotype is applied to this group, suggesting less competence but greater sociability. This is in keeping with the findings of Carlsson and Björklund (2010) with regard to preschool teachers. Our studies are the first to confirm that preservice teachers are subject to negative stereotypes of their competence; in other words, negative stereotypes surrounding the teaching profession are present even during teacher training. We have also shown, for the first time, that stereotype threat has a detrimental effect on the performance of preservice teachers: In our study, the group of preservice teachers who was subject to stereotype threat did worse on a test of cognitive ability than the group who was not. This is, accordingly, the first investigation to show evidence of weaker performance by preservice teachers as soon as stereotype threat is present. The practical implications of our results are discussed below.

It is important to note that these results are of theoretical importance as well. Past research has focused on stereotype threat as it relates to groups in which membership is not chosen and cannot be changed (beyond certain limits), as in the cases of ethnicity and gender. None of the studies cited in the meta-analyses of Nguyen and Ryan (2008) or Walton and Cohen (2003),

Table 3
Study 3: Means (and Standard Deviations) of Test Performance

Variable	Stereotype threat	No stereotype threat
Teacher training	10.69 (2.74)	12.19 (3.88)
Psychology	12.82 (2.61)	12.41 (2.55)

for example, look at groups that people choose to belong to. If membership is voluntary, it is also generally possible to leave the group. As a result, people might regard their membership as less binding and less significant, and this might make them less vulnerable to stereotype threat. However, our studies have shown that stereotype threat has an impact even when membership in a group is freely chosen. The decisive factor is not whether it is possible to leave the group, but whether a negative stereotype of the group exists. Future studies should look more closely at the role of group identification, for example, to find out whether stereotype threat leads to disidentification over the long term, or perhaps even to a departure from the group.

This is an important question for the practical realm as well (i.e., for teacher training). Because even a manipulation that did not explicitly mention the relevant stereotype was sufficient to activate stereotype threat, it appears that the salience of the stereotype, in itself, is enough to produce that result in everyday life. Particularly when preservice teachers attend the same courses as students in other fields, an implicit comparison may well lead them to experience stereotype threat and thus to perform at a lower level. Over the long term, there is the risk of disidentification. Thus, would-be teachers might eventually withdraw from the academic arena or reject an identity that is vulnerable to stereotype threat, instead choosing a different field of study (see Woodcock et al., 2012, among others).

The occurrence of stereotype threat effects on preservice teachers depends on several factors. The general salience of the negative stereotype may not be the same for all countries or cultures. Although we were able to show that there is a negative public perception of teachers and a negative performance-related stereotype of preservice teachers in Germany, the same might not be true in other countries (Alexander, Chant, & Cox, 1994; Everton, Turner, & Hargreaves, 2007; Sahlberg, 2012; Verhoeven, Aelterman, Rots, & Buvens, 2006). In countries or cultures where teaching is a profession of high esteem, teacher training institutions might be able to recruit the very best candidates for teacher training. Under such circumstances, general salience of a negative performance-related stereotype of preservice teachers is unlikely to exist.

Differences between educational systems for teacher training may also influence the occurrence of stereotype threat. In Germany, preservice teachers spend 5 years at the university before they leave the university and start their on-the-job training in schools. During their time at the university, they attend the same courses as students not training to be teachers (e.g., preservice mathematics teachers attend together with students who want to become mathematicians). Here, the preservice teachers face the implicit or explicit comparisons that might lead them to experience stereotype threat. However, during their on-the-job training, preservice teachers are probably safe from stereotype threat for they no longer face comparisons with students from other fields of study and are exposed to positive role models (professional teachers; on the effect of positive role models, see, e.g., Huguet & Régner, 2007). Therefore, with regard to stereotype threat, preservice teachers might benefit from training systems that separate them and other students. Future research may address this point by comparing preservice teachers' vulnerability to stereotype threat (Barnard, Burley, Olivarez, & Crooks, 2008) in different stages of their studies. Additionally, comparisons of stereotype threat effects

on preservice teachers in different teacher training systems might also prove insightful.

Differences within the group of preservice teachers may also be important. We did not investigate this question, owing to the exploratory nature of our studies. It should be noted, however, that preservice teachers differ a great deal with respect to their motivation and cognitive capabilities. Retelsdorf and Möller (2012) investigated motivational predictors of students from different teacher education programs in Germany. They found that subject interest was strongly related to choosing an academic track, whereas educational interest was rather related to the choice of an elementary or nonacademic track program. Kaub and colleagues (2012) found that the profile of those planning to teach science showed a lower level of interest and satisfaction but that their cognitive capacities were superior to those of students planning to teach other subjects. It is also possible that there are different stereotypes of preservice teachers, depending on their specific subject areas. Indeed, research has shown that there are substereotypes within other stereotyped groups (for more on the group of African Americans, for example, see Walzer & Czopp, 2011). Future research on stereotype threat as it relates to preservice teachers should also seek to identify possible substereotypes, as this would paint a more complete picture of how stereotype threat affects this group.

There remain two open questions, which are to be considered in future research. First, Study 3 was not designed to reveal the mechanism of the stereotype threat effect on preservice teachers. Their poor performance might have been a result of cognitive effects such as processes associated with working memory (e.g., Schmader et al., 2008) or motivational effects such as performance avoidance (e.g., Thoman, Smith, Brown, Chase, & Lee, 2013). Considering that any measures taken against stereotype threat effects on preservice teachers will rely on knowledge of their exact mechanism, future research will have to elaborate on that topic. Second, we were not able to consider moderating variables, a well-known factor in stereotype threat research (see, e.g., Martiny & Götz, 2011). Typical examples, which future studies should take into account, include identification with the domain and the stereotyped group. As various researchers have shown (e.g., Keller, 2007; Osborne & Walker, 2006), a high level of identification with a stereotyped domain increases the impact of stereotype threat—although these effects are not entirely clear for all groups or all areas of competence (Nguyen & Ryan, 2008). Paradoxically, preservice teachers who are engaged and identify closely with their field of study might be particularly vulnerable to stereotype threat. The same holds true for identification with the stereotyped group; as Armenta (2010), among others, has shown, this too increases the effect of stereotype threat. One future hypothesis might be that the more someone identifies with becoming a teacher, the greater the potential impact of stereotype threat. Furthermore, preservice teachers' identification with either the group of preservice teachers or their fields of study may depend on their reasons for choosing this profession. This is particularly interesting because the preservice teacher students choose to self-select into a negatively stereotyped profession. It goes beyond the purpose of our study to analyze the motivation for becoming a teacher. However, we know from other research that prospective teachers are strongly motivated to work with children (Paulick, Retelsdorf, & Möller, 2013; Pohlmann & Möller, 2010; Retelsdorf, Bauer, Gebauer, Kauper, &

Möller, in press; Watt et al., 2012) while also being interested in several practical aspects of the teacher profession (e.g., holidays and the good pay [in Germany]). Students joining a teacher training program for intrinsic reasons (e.g., subject interest, educational interest, etc.) might identify stronger with their group or their field of study than students joining for extrinsic reasons (e.g., job security, vacation, etc.). Therefore, students with the more adequate motivation might be the ones most endangered by stereotype threat.

As a final note, we make two methodical observations regarding the realization of any future research. First, future research should make use of more reliable performance measures. The test applied in Study 3 offered only a low (yet still acceptable) internal consistency. The internal consistency found in our study resembles the one offered by the test manual ($\alpha = .66$). The mean score for the participants unaffected by stereotype threat is slightly higher than the standard value specified for participants of that age and educational background, whereas their average standard deviation equals the one given in the test manual. As explained above, the test was chosen for the sake of content, and although a low alpha does not necessarily render a test score useless or impossible to interpret (Carmines & Zeller, 1979; Cronbach, 1951; Schmitt, 1996), the necessity for a more reliable instrument remains. Second, future research should consider a possible mix-up of stereotypes about different social groups that are represented in the sample. As we pointed out in Study 3, other social groups might share a pattern of stereotypes similar to the one encountered for preservice teachers (or any other group that is subject to research, for that matter). Therefore, these social groups might be affected by a stereotype threat manipulation that does not target them. This problem can be dealt with by either excluding possibly problematic groups or, because this is not always possible, by testing for the effects of other group memberships.

These methodical considerations notwithstanding, our studies paint a consistent picture that has clear implications for teacher training. Preservice teachers are viewed as less competent, and this stereotype, when made salient, has a negative impact on their performance. It would therefore be wise to take appropriate measures to counteract that stereotype during teacher training, particularly when students in this group share courses with students in other fields. As a possible first step, instructors and students might be informed of the fact that research has shown that—in contrast to the stereotype—preservice teachers are not, in fact, less competent than other students, but perform at a similar level.

References

Alexander, D., Chant, D., & Cox, B. (1994). What motivates people to become teachers. *Australian Journal of Teacher Education*, 19, 40–49.

Alter, A. L., Aronson, J., Darley, J. M., Rodriguez, C., & Ruble, D. N. (2010). Rising to the threat: Reducing stereotype threat by reframing the threat as a challenge. *Journal of Experimental Social Psychology*, 46, 166–171. doi:10.1016/j.jesp.2009.09.014

Armenta, B. E. (2010). Stereotype boost and stereotype threat effects: The moderating role of ethnic identification. *Cultural Diversity and Ethnic Minority Psychology*, 16, 94–98. doi:10.1037/a0017564

Barnard, L., Burley, H., Olivarez, A., & Crooks, S. (2008). Measuring vulnerability to stereotype threat. *Electronic Journal of Research in Educational Psychology*, 6, 51–64.

Blömeke, S. (2005). Das Lehrerbild in Printmedien. Inhaltsanalyse von "Spiegel"- und "Focus"-Berichten seit 1990 [Teachers in print media: A content analysis of articles of the magazines *Spiegel* and *Focus* since 1990]. *Die Deutsche Schule*, 97, 24–39.

Carlsson, R., & Björklund, F. (2010). Implicit stereotype content: Mixed stereotypes can be measured with implicit association test. *Social Psychology*, 41, 213–222. doi:10.1027/1864-9335/a000029

Carmines, E. G., & Zeller, R. A. (1979). *Reliability and validity assessment*. Beverly Hills, CA: Sage. doi:10.4135/9781412985642

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16, 297–334. doi:10.1007/BF02310555

Davies, P. G., Spencer, S. J., Quinn, D. M., & Gerhardstein, R. (2002). Consuming images: How television commercials that elicit stereotype threat can restrain women academically and professionally. *Personality and Social Psychology Bulletin*, 28, 1615–1628. doi:10.1177/014616702237644

Désert, M., Préaux, M., & Jund, R. (2009). So young and already victims of stereotype threat: Socio-economic status and performance of 6 to 9 years old children on Raven's progressive matrices. *European Journal of Psychology of Education*, 24, 207–218. doi:10.1007/BF03173012

Devine, P. G. (1989). Stereotypes and prejudice: Their automatic and controlled components. *Journal of Personality and Social Psychology*, 56, 5–18. doi:10.1037/0022-3514.56.1.5

Everton, T., Turner, P., & Hargreaves, L. (2007). Public perceptions of the teaching profession. *Research Papers in Education*, 22, 247–265. doi:10.1080/02671520701497548

Fisher, R. J., & Andrews, J. J. (1976). The impact of self-selection and reference group identification in a university living-learning center. *Social Behavior and Personality*, 4, 209–218. doi:10.2224/sbp.1976.4.2.209

Fiske, S. T., Cuddy, A. J. C., Glick, P., & Xu, J. (2002). A model of (often mixed) stereotype content: Competence and warmth respectively follow from perceived status and competition. *Journal of Personality and Social Psychology*, 82, 878–902. doi:10.1037/0022-3514.82.6.878

Fiske, S. T., Xu, J., Cuddy, A. C., & Glick, P. (1999). (Dis)respecting versus (dis)liking: Status and interdependence predict ambivalent stereotypes of competence and warmth. *Journal of Social Issues*, 55, 473–489. doi:10.1111/0022-4537.00128

Huguet, P., & Régner, I. (2007). Stereotype threat among schoolgirls in quasi-ordinary classroom circumstances. *Journal of Educational Psychology*, 99, 545–560. doi:10.1037/0022-0663.99.3.545

Ihme, T. A., & Mauch, M. (2007). Werbung als implizite Aktivierungsquelle von Geschlechterstereotypen und ihr Einfluss auf Mathematikleistungen sowie auf das Computerwissen bei Mädchen und Jungen [Commercials' implicit activation of gender stereotypes and their impact on girls' and boys' performance in mathematics and computer literacy]. *Empirische Pädagogik*, 21, 291–305.

Jordan, A. H., & Lovett, B. J. (2007). Stereotype threat and test performance: A primer for school psychologists. *Journal of School Psychology*, 45, 45–59. doi:10.1016/j.jsp.2006.09.003

Kaub, K., Karbach, J., Biermann, A., Friedrich, A., Bedersdorfer, H., Spinath, F. M., & Brünken, R. (2012). Berufliche Interessensorientierungen und kognitive Leistungsprofile von Lehramtsstudierenden mit unterschiedlichen Fachkombinationen [Vocational interests and cognitive ability of first-year teacher candidates as a function of selected study major]. *Zeitschrift für Pädagogische Psychologie*, 26, 233–249. doi:10.1024/1010-0652/a000074

Keller, J. (2007). Stereotype threat in classroom settings: The interactive effect of domain identification, task difficulty and stereotype threat on female students' maths performance. *British Journal of Educational Psychology*, 77, 323–338. doi:10.1348/000709906X113662

Keller, J., & Dauenheimer, D. (2003). Stereotype threat in the classroom: Dejection mediates the disrupting threat effect on women's math per-

formance. *Personality and Social Psychology Bulletin*, 29, 371–381. doi:10.1177/0146167202250218

Kray, L. J., Thompson, L., & Galinsky, A. D. (2001). Battle of the sexes: Gender stereotype confirmation and reactance in negotiations. *Journal of Personality and Social Psychology*, 80, 942–958. doi:10.1037/0022-3514.80.6.942

Liepmann, D., Beauducel, A., Brocke, B., & Amthauer, R. (2007). *Intelligenz-Struktur-Test 2000R* [Intelligence Structure Test 2000R]. Göttingen, Germany: Hogrefe.

Martiny, S. E., & Götz, T. (2011). Stereotype Threat in Lern- und Leistungssituationen: Theoretische Ansätze, empirische Befunde und praktische Implikationen [Stereotype threat in learning and achievement situations: Theoretical approaches, empirical results, and practical implications]. In M. Dresel & L. Lämmle (Eds.), *Motivation, Selbstregulation und Leistungsexzellenz* (Talentförderung – Expertiseentwicklung – Leistungsexzellenz, Vol. 9, pp. 153–177). Münster, Germany: LIT.

McKown, C., & Weinstein, R. S. (2003). The development and consequences of stereotype consciousness in middle childhood. *Child Development*, 74, 498–515. doi:10.1111/1467-8624.7402012

Milner, H. R., & Woolfolk Hoy, A. (2003). A case study of an African American teacher's self-efficacy, stereotype threat, and persistence. *Teaching and Teacher Education*, 19, 263–276. doi:10.1016/S0742-051X(02)00099-9

Muzzatti, B., & Agnoli, F. (2007). Gender and mathematics: Attitudes and stereotype threat susceptibility in Italian children. *Developmental Psychology*, 43, 747–759. doi:10.1037/0012-1649.43.3.747

Nguyen, H.-H. D., & Ryan, A. M. (2008). Does stereotype threat affect test performance of minorities and women? A meta-analysis of experimental evidence. *Journal of Applied Psychology*, 93, 1314–1334. doi:10.1037/a0012702

Osborne, J. W., & Walker, C. (2006). Stereotype threat, identification with academics, and withdrawal from school: Why the most successful students of colour might be the most likely to withdraw. *Educational Psychology: An International Journal of Experimental Educational Psychology*, 26, 563–577. doi:10.1080/01443410500342518

Paulick, I., Retelsdorf, J., & Möller, J. (2013). Motivation for choosing teacher education: Relations with teachers' achievement goals and instructional practices. *International Journal of Educational Research*, 61, 60–70. doi:10.1016/j.ijer.2013.04.001

Pohlmann, B., & Möller, J. (2010). Fragebogen zur Erfassung der Motivation für die Wahl des Lehramtsstudiums (FEMOLA) [Motivation for choosing teacher education questionnaire (FEMOLA)]. *Zeitschrift für Pädagogische Psychologie*, 24, 73–84. doi:10.1024/1010-0652/a000005

Reips, U.-D. (2002). Standards for internet-based experimenting. *Experimental Psychology*, 49, 243–256. doi:10.1027//1618-3169.49.4.243

Retelsdorf, J., Bauer, J., Gebauer, S. K., Kauper, T., & Möller, J. (in press). Erfassung berufsbezogener Selbstkonzepte von angehenden Lehrkräften (ERBSE-L) [Measuring prospective teachers' professional self-concept]. *Diagnostica*.

Retelsdorf, J., & Möller, J. (2012). Grundschule oder Gymnasium? Zur Motivation ein Lehramt zu studieren [Primary or secondary school? On the motivation of choosing teacher education]. *Zeitschrift für Pädagogische Psychologie*, 26, 5–17. doi:10.1024/1010-0652/a000056

Rydell, R. J., Rydell, M. T., & Boucher, K. L. (2010). The effect of negative performance stereotypes on learning. *Journal of Personality and Social Psychology*, 99, 883–896. doi:10.1037/a0021139

Rydell, R. J., Shiffrin, R. M., Boucher, K. L., Van Loo, K., & Rydell, M. T. (2010). Stereotype threat prevents perceptual learning. *Psychological and Cognitive Sciences*, 107, 14042–14047. doi:10.1073/pnas.1002815107

Sahlberg, P. (2012). The most wanted: Teachers and teacher education in Finland. In A. Lieberman & L. Darling-Hammond (Eds.), *Teacher education around the world: Changing policies and practices* (pp. 1–21). New York, NY: Routledge.

Schmader, T., Johns, M., & Forbes, C. (2008). An integrated process model of stereotype threat effects on performance. *Psychological Review*, 115, 336–356. doi:10.1037/0033-295X.115.2.336

Schmitt, N. (1996). Uses and abuses of coefficient alpha. *Psychological Assessment*, 8, 350–353. doi:10.1037/1040-3590.8.4.350

Spinath, B., van Ophuysen, S., & Heise, E. (2005). Individuelle Voraussetzungen von Studierenden zu Studienbeginn: Sind Lehramtsstudierende so schlecht wie ihr Ruf? [University students' learning- and achievement-related characteristics: The case of teacher students]. *Psychologie in Erziehung und Unterricht*, 52, 186–197.

Steele, C. M. (1997). A threat in the air: How stereotypes shape intellectual identity and performance. *American Psychologist*, 52, 613–629. doi:10.1037/0003-066X.52.6.613

Steele, C. M. (2010). *Whistling Vivaldi and other clues to how stereotypes affect us*. New York, NY: Norton.

Steele, C. M., & Aronson, J. (1995). Stereotype threat and the intellectual test performance of African Americans. *Journal of Personality and Social Psychology*, 69, 797–811. doi:10.1037/0022-3514.69.5.797

Swetnam, L. A. (1992). Media distortion of the teacher image. *Clearing House*, 66, 30–32. doi:10.1080/00098655.1992.9955921

Taylor, V. J., & Walton, G. M. (2011). Stereotype threat undermines academic learning. *Personality and Social Psychology Bulletin*, 37, 1055–1067. doi:10.1177/0146167211406506

teaching. (2013). In *Encyclopædia Britannica*. Retrieved from <http://www.britannica.com/EBchecked/topic/585183/teaching>

Thoman, D. B., Smith, J. L., Brown, E. R., Chase, J., & Lee, J. Y. K. (2013). Beyond performance: A motivational experiences model of stereotype threat. *Educational Psychology Review*, 25, 211–243. doi:10.1007/s10648-013-9219-1

Verhoeven, J. C., Aelterman, A., Rots, I., & Buven, I. (2006). Public perceptions of teachers' status in Flanders. *Teachers and Teaching: Theory and Practice*, 12, 479–500. doi:10.1080/13450600600644350

von Hippel, C., Kalokerinos, E. K., & Henry, J. D. (2013). Stereotype threat among older employees: Relationship with job attitudes and turnover intentions. *Psychology and Aging*, 28, 17–27. doi:10.1037/a0029825

Walton, G. M., & Cohen, G. L. (2003). Stereotype lift. *Journal of Experimental Social Psychology*, 39, 456–467. doi:10.1016/S0022-1031(03)00019-2

Walzer, A. S., & Czopp, A. M. (2011). Able but unintelligent: Including positively stereotyped black subgroups in the stereotype content model. *Journal of Social Psychology*, 151, 527–530. doi:10.1080/00224545.2010.503250

Watt, H. M. G., Richardson, P. W., Klusmann, U., Kunter, M., Beyer, B., Trautwein, U., & Baumert, J. (2012). Motivations for choosing teaching as a career: An international comparison using the FIT-Choice scale. *Teaching and Teacher Education*, 28, 791–805. doi:10.1016/j.tate.2012.03.003

Woodcock, A., Hernandez, P. R., Estrada, M., & Schultz, P. W. (2012). The consequences of chronic stereotype threat: Domain disidentification and abandonment. *Journal of Personality and Social Psychology*, 103, 635–646. doi:10.1037/a0029120

Received September 24, 2013

Revision received June 3, 2014

Accepted June 5, 2014 ■

Value Development Underlies the Benefits of Parents' Involvement in Children's Learning: A Longitudinal Investigation in the United States and China

Cecilia Sin-Sze Cheung
University of California, Riverside

Eva M. Pomerantz
University of Illinois, Urbana-Champaign

This research examined whether the benefits of parents' involvement in children's learning are due in part to value development among children. Four times over the 7th and 8th grades, 825 American and Chinese children (M age = 12.73 years) reported on their parents' involvement in their learning and their perceptions of the value their parents place on school achievement as well as the value they themselves place on it. Children's academic functioning was assessed via children's reports and school records. Value development partially explained the effects of parents' involvement on children's academic functioning in the United States and China. For example, the more children reported their parents as involved, the more they perceived them as placing value on achievement 6 months later; such perceptions in turn predicted the subsequent value children placed on achievement, which foreshadowed enhanced grades.

Keywords: achievement, engagement, parent involvement, socialization, value transmission

A wealth of research supports the idea that parents' involvement in children's learning enhances children's academic functioning (for reviews, see Grolnick, Friendly, & Bellas, 2009; Pomerantz, Kim, & Cheung, 2012). Children whose parents are involved on the school (e.g., attending parent-teacher conferences) and home (e.g., discussing school with children) fronts often exhibit enhanced engagement (e.g., use of self-regulated strategies), skills (e.g., phonological awareness), and achievement (e.g., grades). Notably, parents' involvement plays a role in children's academic functioning even when aspects of children's home environment such as parents' income and education are taken into account (e.g., Dearing, Kreider, Simpkins, Weiss, 2006; Jeunes, 2005, 2007). The effects of parents' involvement are also not accounted for by other dimensions of parenting such as supporting children's autonomy (e.g., C. S. Cheung & Pomerantz, 2011; Deslandes, Bouchard, & St.-Amant, 1998).

Research focusing on why parents' involvement in children's learning benefits children's academic functioning identifies the

development of children's actual and perceived competencies as important (e.g., Dearing et al., 2006; Senechal & LeFevre, 2002). However, it has also been argued that parents' involvement leads children to view doing well in school as valuable, which fosters children's engagement in school, enhancing their achievement (e.g., Epstein, 1988; Grolnick & Slomiaczek, 1994). Unfortunately, such a value development model has not been tested. The goal of the current research was to address this gap by evaluating whether the effect of parents' involvement on children's engagement and grades in school is due in part to the development of children's values in regard to school achievement. Drawing on prior theory and research on value transmission (i.e., children's adoption of parents' values; e.g., Grusec & Goodnow, 1994; Knafo & Schwartz, 2009) as well as parents' involvement in children's learning (e.g., Hill & Tyson, 2009), we hypothesized two pathways by which parents' involvement facilitates children valuing achievement in school.

The Perception-Acceptance Value Development Pathway

Grusec and Goodnow (1994) proposed a two-step process model by which parents transmit their values to children. First, children must be aware of parents' values such that they perceive them accurately. Second, children must accept parents' values as their own. Both steps are considered key in effective transmission of values from generation to generation (e.g., Barni, Ranieri, Scabini, & Rosnati, 2011; Knafo & Schwartz, 2009). Grusec and Goodnow focused on how the type of discipline parents use with children contributes to value transmission by shaping children's perceptions of parents' values. Several other dimensions of interactions between parents and children, such as parents' discussion of their values with children (e.g., Knafo & Schwartz, 2004;

This article was published Online First July 28, 2014.

Cecilia Sin-Sze Cheung, Department of Psychology, University of California, Riverside; Eva M. Pomerantz, Department of Psychology, University of Illinois, Urbana-Champaign.

This research was supported by National Institute of Mental Health Grant R01 MH57505. We appreciate the constructive comments on an earlier version of this article provided by members of the Center for Parent Child Studies at University of Illinois, Urbana-Champaign.

Correspondence concerning this article should be addressed to Cecilia Sin-Sze Cheung, Department of Psychology, University of California, Riverside, 900 University Avenue, Riverside, CA 92521, or Eva M. Pomerantz, Department of Psychology, University of Illinois, Urbana-Champaign, 603 East Daniel Street, Champaign, IL 61820. E-mail: ccheung@ucr.edu or pomerantz@illinois.edu

Okagaki & Bevis, 1999) and the quality of children's relationships with parents (Barni et al., 2011), have also received attention. As a commitment of resources (e.g., time, energy, and financial provisions) to children in the academic arena (Grolnick & Slawiaczek, 1994), parents' involvement in children's learning may be a key mechanism by which parents convey to children that they view school as important. When parents take the time and trouble to participate in school events, children may view parents as placing importance on learning. Parents' involvement on the home front may have similar consequences—for example, when parents ask children about what they are learning in school or provide children with learning resources (e.g., books), they may communicate that they see doing well in school as useful.

When children see parents as valuing achievement in school, they may come to value it themselves (e.g., Eccles et al., 1983; Grolnick, Ryan, & Deci, 1997). Grusec and Goodnow (1994) argued that once children are aware of the values parents hold their acceptance of such values as their own is facilitated in the context of a warm relationship with parents (see also Barni et al., 2011). The commitment of resources characteristic of parents' involvement may signal to children that parents care about them. Moreover, in the context of their involvement, parents may provide emotional support for children (e.g., by reacting to children's frustration with homework with soothing words), thereby creating a sense of trust in children that may facilitate their adoption of parents' values (e.g., C. S. Cheung & Pomerantz, 2012; Grolnick & Slawiaczek, 1994; Grusec, 2002). Parents' involvement in children's learning may be a particularly unique dimension of parenting in that it simultaneously communicates the value parents place on doing well in school (Step 1 of Grusec and Goodnow's model), while also leading children to take on this value as their own (Step 2 of Grusec and Goodnow's model). Thus, parents' involvement may enhance children's achievement via a *perception-acceptance pathway*: Parents' involvement leads children to perceive parents as valuing school achievement (path a in Figure 1), thereby heightening the value children themselves place on it (path b in Figure 1).

The Experience Value Development Pathway

The perception-acceptance pathway may be accompanied by what we label an experience pathway that directly fosters the value children place on achievement in school. Although parents' involvement in children's learning likely conveys the value parents' place on children's school endeavors, it may not always lead children to value school achievement via children's awareness of parents' values (Eccles et al., 1983). When parents become involved in children's learning, they may create experiences for children that directly heighten the value children place on school

achievement. For example, when parents discuss school with children, children may generate reasons for its utility, leading them to see doing well in school as valuable (Hill & Tyson, 2009). In a somewhat different vein, drawing from Bem's (1967, 1972) Self-Perception Theory, practices such as helping children to sustain their effort on their homework until it is finished may lead children to conclude that they value doing well in school given how much time they invest in it. In the *experience pathway*, parents' involvement creates experiences that lead children to place value on school achievement (path c in Figure 1), regardless of their perceptions of parents' values.

The Role of Values in Academic Functioning

Whether the value children place on achievement in school ensuing from parents' involvement develops via a perception-acceptance pathway or an experience pathway, prior theory and research (e.g., Eccles et al., 1983; Wang & Pomerantz, 2009) indicates it supports children's academic functioning (see path d in Figure 1). In their Expectancy-Value Theory, Eccles et al. (1983) made the case that when children value achievement in school, they become more engaged in school, which enhances their achievement. Indeed, the more children view doing well in school as important, the more engaged they are—for example, they use heightened self-regulated learning strategies, such as monitoring and planning their learning (e.g., Pintrich, 1999; Wang & Pomerantz, 2009). Notably, heightened value as well as engagement predicts improved achievement among children over time (e.g., Alexander, Entwistle, & Dauber, 1993; Kenney-Benson, Pomerantz, Ryan, & Patrick, 2006; Wang & Pomerantz, 2009).

Value Development Pathways in the United States and China

Over the last several years, there has been a call to extend the understanding of psychological processes beyond Western populations (e.g., Arnett, 2008; Henrich, Heine, Norenzayan, 2010a, 2010b). In the case of parents' involvement in children's learning, this may be of import when it comes to China because Chinese parents are involved differently in children's learning than are their American counterparts (for a review, see Pomerantz, Ng, Cheung, & Qu, in press). For one, Chinese (vs. American) parents are more involved compared to American parents (e.g., Chen & Stevenson, 1989; Ng, Pomerantz, & Lam, 2007). Consequently, both the perception-acceptance and experience value development pathways may be stronger in China than the United States as such heightened involvement may convey more clearly that parents value school achievement and create more experiences that di-

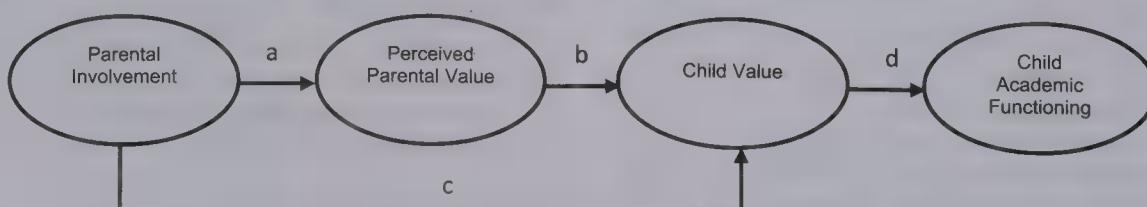


Figure 1. Hypothesized value development pathways underlying the effect of parental involvement on children's academic functioning. The perception-acceptance pathway is reflected in paths a, b, and d; the experience pathway is reflected in paths c and d.

rectly heighten the value children place on school achievement. Moreover, the amplified commitment of resources reflected in parents' involvement may enhance children's adoption of parents' values.

Chinese parents' involvement in children's learning, however, is more controlling than that of American parents with greater attention to children's mistakes (e.g., C. S. Cheung & Pomerantz, 2011; Ng et al., 2007). This along with the tendency for Chinese (vs. American) children to feel less close to parents during adolescence (e.g., Pomerantz, Qin, Wang, & Chen, 2009) may undermine value transmission. Although it is unclear if parents' involvement similarly fosters value development in China and United States, prior examination of the effects of parents' involvement on children's engagement and grades yields similar effects in the two countries (C. S. Cheung & Pomerantz, 2011).

Overview of the Current Research

To examine whether parents' involvement in children's learning enhances children's academic functioning by heightening the value children place on school achievement in the United States and China, the current research evaluated the hypothesis that two value-development pathways underlie the benefits of parents' involvement (see Figure 1). In the *perception-acceptance pathway* (paths a, b, and d), parents' involvement signals to children that parents value school achievement, leading children to value it, which in turn enhances children's achievement. In the *experience*

pathway (paths c and d), parents' involvement develops the value children place on school achievement not through the messages it conveys about parents' values, but rather directly through the experiences it creates. Comparisons between the United States and China for both pathways were made to evaluate their generalizability.

In testing the value transmission pathways, we focused on children in the middle school years because parents' involvement may offset the devaluing of school that often occurs among children during this phase of development (for a review, see Wigfield & Wagner, 2005). Children in the United States and China reported four times over the seventh and eighth grades on parents' involvement in their learning, their perceptions of the value parents place on school achievement, and the value they themselves place on it. Children's academic functioning was assessed with children's reports and school records. The four-wave design allowed for the examination of the sequence of effects posited in Figure 1. Because each construct was assessed at each wave, autoregressive effects could be taken into account (see Figure 2), which permitted identification of the direction of effects.

We investigated two dimensions of children's academic functioning that have important implications for children's lives. First, children's *engagement* in school is not only predictive of their achievement over time (e.g., M.-T. Wang & Fredricks, 2014; Q. Wang & Pomerantz, 2009) but also appears to protect children against internalizing and externalizing problems (e.g., M.-T. Wang

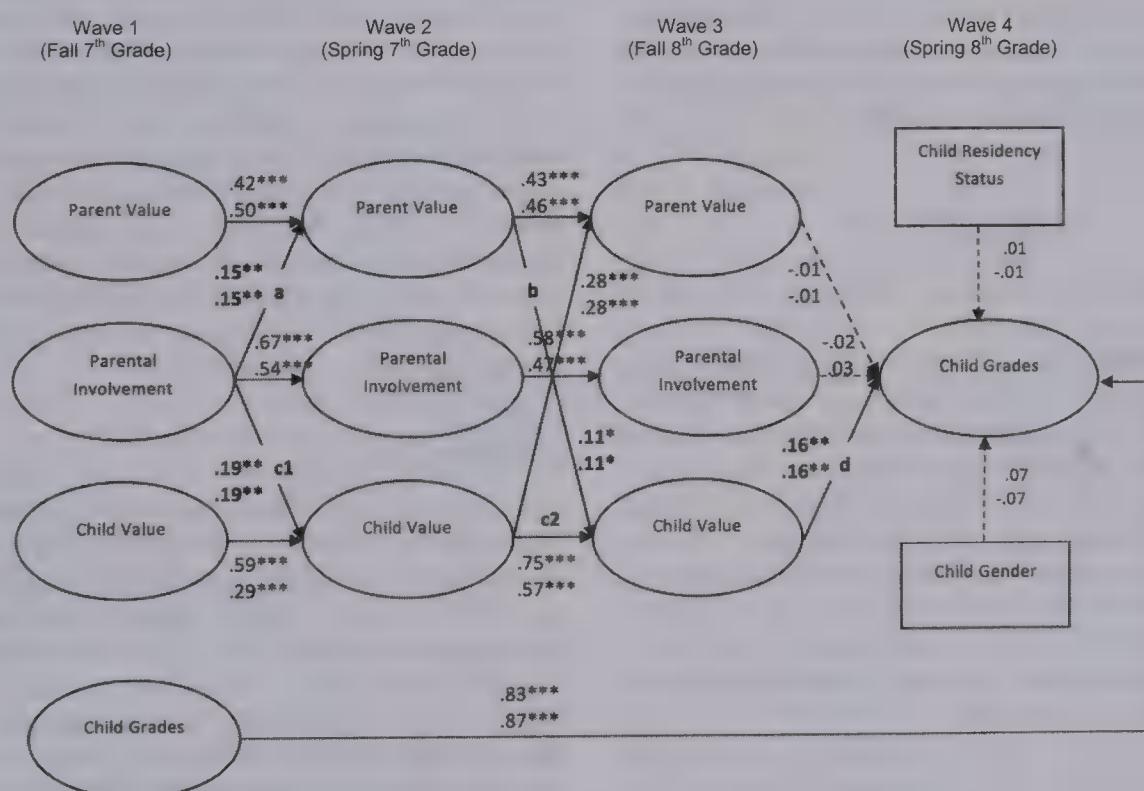


Figure 2. Value development pathways underlying the effect of parents' involvement on children's grades. For child gender, 1 = boys, 2 = girls; for child residence with parents, 1 = not residing with both parents, 2 = residing with both parents. Letters (i.e., a, b, c1, c2, and d) represent links comprising the two value development pathways denoted in Figure 1. For ease of presentation, within-wave covariances are not shown. Based on the chi-square difference tests, all paths comprising the indirect pathways were constrained to be equal between the United States and China. American standardized estimates are above; Chinese standardized estimates are below. Solid lines are significant ($p < .05$); dashed lines are not. * $p < .05$. ** $p < .01$. *** $p < .001$.

& Fredricks, 2014; M.-T. Wang & Peck, 2013). Children reported on two forms of their engagement—their use of self-regulated learning strategies and the time they spend on schoolwork outside of school. Second, children's *grades* in school are a significant reflection of their achievement (Duckworth & Seligman, 2005; Grolnick et al., 1997) with implications for subsequent opportunities (e.g., placement in enrichment activities) as well as success later in life (e.g., Geiser & Santelices, 2007). There is sizeable evidence documenting the importance of parents' involvement in children's learning for both children's engagement and grades (e.g., C. S. Cheung & Pomerantz, 2011; Grolnick & Slowiaczek, 1994).

With the exception of grades, children provided reports for all the constructs under study. This is of particular concern when it comes to parents' involvement in children's learning. Children's reports of such involvement are only modestly associated with teachers and parents' reports (e.g., Bakker, Denessen, & Brus-Laeven, 2007; Hill et al., 2004; Reynolds, 1992). However, because children, teachers, and parents' reports of parents' involvement each predict unique variance in children's achievement, it has been argued that each captures unique aspects of parents' involvement (Reynolds, 1992). Children's reports reflect their *perceptions* of parents' involvement. This is significant because children must notice parents' involvement to draw conclusions about parents' values (C. S. Cheung & Pomerantz, 2012; Grolnick & Slowiaczek, 1994). However, each reporter may also bring a unique set of biases to their reports. In the current context, the value children view parents placing on school achievement or that they themselves place on it may bias their reports, such that effects reflect children's perceptions of parents' values or their own values rather than parents' involvement. To rule out this possibility, we tested alternative pathways—for example, the value children place on school achievement predicts their reports of parents' involvement over time.

Method

Participants

The University of Illinois U.S.-China Adolescence Study began when children entered a new school in seventh grade and concluded at the end of eighth grade in the United States and China (e.g., Pomerantz et al., 2009; Wang & Pomerantz, 2009). Participants were 374 American children (187 boys; M age = 12.78 years in the fall of seventh grade) and 451 Chinese children (240 boys; M age = 12.69 years in the fall of seventh grade). In each country, children attended public school in primarily working- or middle-class areas. The American children attended one of two public schools consisting of the seventh and eighth grades in the suburbs of Chicago. Chicago is a city with high population density (12,750 people per square mile at the time of the research) with a median yearly family gross income of \$61,182 at the time of the research; 30% of the population over the age of 25 possessed at least a college degree at the time of the research (U.S. Census Bureau, 2007). The median family income of the two selected suburbs was \$60,057 and \$72,947, with 21% and 26% of the population over the age of 25 possessing a college degree. Reflecting the ethnic composition of these areas, participants were predominantly European American (88%) with 9% Hispanic American, 2% African

American, and 1% Asian American. Seventy-nine percent of participating children reported living with two parents.

The Chinese children attended one of two public schools in the suburbs of Beijing; one school consisted of the seventh to ninth grades and the other of the seventh to 12th grades. According to the Beijing Municipal Bureau of Statistics (2005), Beijing is a densely populated city (13,386 people per square mile at the time of the research) with an annual discretionary income per capita of \$15,638 RMB at the time of the research; 13% of the population over the age of 6 had at least a college degree at the time of the research. In the two selected suburbs, 9% and 28% of the population over the age of 6 had a college degree. Over 95% of the residents in these areas were of the *Han* ethnicity (Beijing Municipal Bureau of Statistics, 2005), which is slightly above the 92% for the country as a whole (China Population and Development Research Center, 2001). Eighty-six percent of the participating children reported living with two parents. An opt-in consent procedure was used in which parents provided permission for children to participate. Sixty-four percent of parents in the United States and 59% of parents in China allowed their children to participate.

Procedure

Children completed a set of questionnaires during two 45-min sessions at four times approximately 6 months apart: fall of seventh grade (Wave 1), spring of seventh grade (Wave 2), fall of eighth grade (Wave 3), and spring of eighth grade (Wave 4). Instructions and items were read aloud to children in their native language in the classroom during regular class time by trained native research staff. Children received a small gift (e.g., a calculator) as a token of appreciation at the end of each session. The average attrition rate over the entire study was 4% (2% in the United States and 6% in China). More than 85% of the children had data at all four waves of the study for all of the analyses, with more than 98% having data at two or more waves for all of the analyses. At Wave 1, children with complete data differed from those without complete data only in that their grades were better, $t(818) = 2.01$, $p < .05$. The Institutional Review Boards of the University of Illinois and Beijing Normal University approved the procedures.

Measures

The measures were originally written in English. Standard translation and back-translation procedures (Brislin, 1980) were employed with repeated discussion among American and Chinese members of the research team to modify the wording of the items to ensure equivalence in meaning between the English and Chinese versions (Erkut, 2010). Equivalence was also established statistically. A series of confirmatory factor analyses (CFAs) was conducted in the context of two-group nested structural equation modeling (SEM) to examine the metric invariance of the measures between the United States and China over the four waves of the study; metric invariance is essential and sufficient in making valid comparisons of the associations (e.g., Little, 1997), as was done in the current research (see below).

In each set of CFAs, an unconstrained model was compared to a constrained (i.e., metric invariance) model. The unconstrained models consisted of the same latent construct repeatedly assessed

over the four waves yielding a total of four latent constructs. These constructs were allowed to correlate with one another; errors of the same indicators over time were also allowed to correlate when suggested by modification indexes from the CFAs conducted on the sample with no missing data (Keith, 2006; McDonald & Ho, 2002). The parameters in the unconstrained models were freely estimated without any between-country or across-time equality constraints. In the constrained models, the factor loadings of the same indicators were forced to be equal between the two countries and across the four waves. Monte Carlo studies indicate that a decrease from the unconstrained to the corresponding constrained model in the comparative fit index (CFI) of no more than .01, supplemented by an increase in the root-mean-square error of approximation (RMSEA) of no more than .015, is reflective of invariance (Chen, 2007). Although chi-square difference tests are considered appropriate for hypothesis testing purposes, the current consensus is that they are not appropriate for evaluating measurement invariance (e.g., Chen, 2007; G. W. Cheung & Rensvold, 2002; Little, 1997).

Prior analyses on these data, using two parcels of items (or two items in the case of time spent on homework outside of school) to represent each latent construct indicated that the measures of parents' involvement in children's learning, the value children place on school, and children's engagement have metric invariance between countries and over time (C. S. Cheung & Pomerantz, 2011; Wang & Pomerantz, 2009). The use of parcels allowed us to build parsimonious models based on solid and meaningful indicators, enhancing the likelihood of replication in future research (Little, Cunningham, Shahar, & Widaman, 2002; Little, Rhemtulla, Gibson, & Schoemann, 2013). Parsimony was of particular concern in the current research given the sizeable number of items comprising each scale and the complexity of the models, which can strain the number of free parameters that can be estimated (e.g., Kline, 1998), despite our sample size of 825. In such a case, the use of parcels is desirable (Little et al., 2013). Importantly, principal components analysis (PCA) on each set of items comprising each parcel indicated that each set formed a single factor; the parcels were also each internally reliable on their own ($\alpha_s = .73$ to $.88$).

Metric invariance of children's perceptions of the value parents place on school achievement has not been evaluated in prior analyses; thus, it was tested for the current research. The latent construct was represented by two parcels of items: Items about the importance of doing well were aggregated in one parcel, which PCA indicated formed a single factor ($\alpha_s = .80$ to $.93$; see item descriptions below), and items about the importance of not doing poorly were aggregated in another, which PCA indicated form a single factor as well ($\alpha_s = .84$ to $.92$). Both the unconstrained, $\chi^2(df = 9) = 26.10$, CFI = $.95$, Tucker-Lewis index (TLI) = $.92$, RMSEA = $.08$, and constrained, $\chi^2(df = 13) = 38.12$, CFI = $.95$, TLI = $.92$, RMSEA = $.07$, models fit the data adequately, with differences between the CFIs and RMSEAs of no more than .01.

Parental involvement in child learning. Parents' involvement in children's learning was assessed with 10 items (e.g., "My parents help me with my homework when I ask." "My parents try to get to know the teachers at my school." "My parents purchase extra workbooks or outside materials related to school for me.") adapted from prior research (Chao, 2000; Kerr & Stattin, 2000; Kohl, Lengua, McMahon, & The Conduct Problems Prevention

Research Group, 2000; Stattin & Kerr, 2000). In line with Grodnick and Slowiaczek's (1994) definition of parents' involvement, the items characterize a variety of practices (e.g., attendance of parent-teacher conferences, discussion of school with children, and assistance with homework) reflecting parents' commitment of resources to children in the academic arena. Children indicated the extent to which each of the statements was true (1 = *not at all true*, 5 = *very true*). The 10 items were combined, with higher numbers reflecting greater involvement as reported by children ($\alpha_s = .83$ to $.85$ in the United States and $.77$ to $.83$ in China).

Child perceptions of parental value. To assess children's perceptions of the value their parents place on school achievement, children indicated how important (1 = *not at all important*, 7 = *very important*) it is to parents that they do well (e.g., "How important is it to your parents that you do well in language arts?") and avoid doing poorly (e.g., "How important is it to your parents that you avoid doing poorly in math?") on four core subjects (language arts, math, science, and social studies in the United States; language arts, math, biology, and English in China) for which children received grades. The eight items were combined, with higher numbers reflecting perceptions of greater parental value ($\alpha_s = .93$ to $.96$ in the United States and $.87$ to $.91$ in China).

Child value. The value children themselves place on school achievement was assessed with a modified version of Pomerantz, Saxon, and Oishi's (2000) measure. Paralleling the measure of children's perceptions of the value parents place on school achievement, for each of the four core subjects, children indicated how important (1 = *not at all important*, 7 = *very important*) it was for them to do well (e.g., "How important is it to you to do well in math?") and avoid doing poorly (e.g., "How important is it to you to avoid doing poorly in language arts?"). The eight items were combined, with higher numbers reflecting greater value ($\alpha_s = .91$ to $.94$ in the United States and $.88$ to $.91$ in China).

Child engagement. Two forms of children's engagement in school were assessed. The 30-item Dowson and McInerney (2004) Goal Orientation and Learning Strategies Survey assessed children's use of *self-regulated learning strategies*. Three scales assess children's metacognitive strategies: Six items assess monitoring (e.g., "I check to see if I understand the things I am trying to learn"), six assess planning (e.g., "I try to plan out my schoolwork as best as I can"), and six assess regulating (e.g., "If I get confused about something at school, I go back and try to figure it out"). Two scales assess children's cognitive strategies: Six items assess rehearsal (e.g., "When I want to learn things for school, I practice repeating them to myself") and six assess elaboration (e.g., "I try to understand how the things I learn in school fit together with each other"). Children indicated the extent to which each of the 30 statements was true of them (1 = *not at all true*, 5 = *very true*). The metacognitive and cognitive strategies scales were combined, with higher numbers representing greater school engagement ($\alpha_s = .96$ to $.97$ in the United States and $.93$ to $.96$ in China).

The *time children spend on schoolwork outside of school* was assessed with a modified version of the scale used by Fuligni, Tseng, and Lam (1999). Children indicated how much time they spend on their schoolwork outside of school on a typical weekday and weekend (1 = *less than 1 hr*, 6 = *more than 5 hr*). Their responses for a typical weekday were weighted by five and combined with those of each day for a typical weekend weighted by two. Higher numbers reflect more time spent on schoolwork out-

side of school ($rs = .48$ to $.64$ in the United States and $.41$ to $.52$ in China).

Child grades. Children's grades in the four core subjects were obtained from schools. Grades in the American schools were originally in letters and were converted to numbers. Because there were 13 steps in the ladder of grades used in the American schools, grades were converted to numbers with a range of 0 (i.e., a grade of F) to 12 (i.e., a grade of A+) with a 1-point increment between each step in the grades (e.g., B- = 7, B = 8, B+ = 9, A- = 10). Such conversion has been used in prior research (e.g., Coe, Pivarnik, Womack, Reeves, & Malina, 2006; Schwartz, Kelly, & Duong, 2013; Wood & Locke, 1987). Moreover, simulation research indicates that the treatment of discrete categories as continuous is unlikely to result in biased parameter estimates when the number of categories is more than six as is the case in the current research (Rhemtulla, Brosseau-Liard, & Savalei, 2012). In the Chinese schools, grades were originally numerical, ranging from 0 to 100 in one school and from 0 to 120 in the other. In both countries, grades were standardized within school to take into account differences in the grading systems of the schools. The four subjects were combined, with higher numbers reflecting better grades.

Results

Overall, the measures in the current research were approximately normally distributed. In both the United States and China across the four waves of assessment, the indexes for skewness and kurtosis were less than 1, with only one exception—the index for

skewness was 1.47 and the kurtosis index was 2.28 for the avoidant dimension of the value children place on school achievement at Wave 1 in the United States. Hence, across the six measures at each of the four waves there was no indication of serious violation of the normality assumption.

As shown in Table 1, in both the United States and China, parents' involvement in children's learning—as reported by children—was positively associated with children's perceptions of the value parents place on school achievement ($rs = .25$ to $.43$, $ps < .001$) as well as the value children themselves place on it ($rs = .25$ to $.39$, $ps < .001$) at each wave. Children's perceptions of the value parents place on school achievement were positively associated at each wave with the value children place on it ($rs = .28$ to $.55$, $ps < .001$). The value children place on school achievement was also associated with their engagement ($rs = .30$ to $.58$ for self-regulated learning strategies and $.13$ to $.21$ for time spent on school schoolwork outside of school; $ps < .05$) as well as grades ($rs = .22$ to $.38$, $ps < .001$) at each wave in the United States and China. Although such associations are suggestive of the viability of both the perception-acceptance and experience value development pathways, they do not provide insight into the direction of effects. Evaluation of the direction of effects requires analyses accounting for the autoregressive effects.

The central analyses took such effects into account. These analyses were conducted within a latent SEM framework using Mplus 7.0 (Muthén & Muthén, 1998–2012), which employs full information maximum likelihood (FIML) estimation in the presence of missing data; FIML provides more reliable standard errors

Table 1
Means and Correlations Among the Central Constructs

Variable	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Parental involvement																
1. Wave 1	—	.50	.54	.46	.25	.24	.25	.27	.27	.24	.18	.19	.01	.04	.07	.10
2. Wave 2	.64	—	.59	.55	.26	.35	.29	.26	.26	.32	.26	.28	.04	.05	.06	.06
3. Wave 3	.51	.60	—	.63	.15	.27	.34	.26	.16	.28	.25	.27	.07	.10	.11	.12
4. Wave 4	.48	.45	.48	—	.15	.29	.26	.33	.12	.26	.24	.27	.05	.05	.09	.11
Perceived parental value																
5. Wave 1	.31	.27	.21	.28	—	.39	.38	.34	.51	.28	.28	.21	.05	.03	.10	.12
6. Wave 2	.34	.43	.27	.22	.41	—	.53	.48	.32	.43	.39	.41	.11	.12	.16	.15
7. Wave 3	.30	.32	.38	.29	.39	.52	—	.57	.33	.44	.55	.50	.12	.16	.17	.17
8. Wave 4	.21	.16	.24	.38	.34	.36	.54	—	.31	.40	.46	.49	.09	.12	.12	.13
Child value																
9. Wave 1	.38	.30	.28	.21	.28	.28	.33	.24	—	.49	.43	.37	.28	.26	.24	.30
10. Wave 2	.44	.39	.36	.26	.28	.41	.43	.30	.67	—	.61	.59	.36	.38	.31	.39
11. Wave 3	.38	.33	.35	.27	.22	.36	.47	.38	.58	.73	—	.67	.27	.33	.29	.31
12. Wave 4	.31	.32	.34	.33	.20	.31	.46	.41	.55	.65	.75	—	.27	.28	.27	.32
Grades																
13. Wave 1	.09	.13	.22	.14	-.03	.01	.12	.13	.28	.20	.20	.31	—	.91	.82	.88
14. Wave 2	.12	.15	.22	.16	-.01	.05	.11	.13	.28	.22	.27	.33	.92	—	.86	.91
15. Wave 3	.11	.15	.22	.15	.03	.07	.14	.12	.26	.24	.31	.35	.82	.87	—	.88
16. Wave 4	.12	.15	.20	.16	.00	.10	.13	.16	.30	.27	.35	.38	.80	.84	.92	—
Mean (U.S.)	3.61	3.44	3.43	3.37	6.38	6.14	6.09	6.10	5.65	5.29	5.35	5.29				
SD (U.S.)	0.71	0.80	0.76	0.76	0.79	1.07	1.05	1.11	1.10	1.18	1.25	1.29				
Mean (China)	3.79	3.69	3.67	3.64	6.00	6.10	6.08	6.05	5.91	5.82	5.78	5.74				
SD (China)	0.62	0.71	0.68	0.68	0.94	0.92	0.94	0.94	0.88	1.08	1.08	1.13				

Note. Results are based on the observed, rather than latent, variables. Correlations for the American sample are presented in the lower triangle; those for the Chinese sample are presented in the upper triangle. Correlations with absolute values greater than .10 are significant ($p < .05$). Grades were standardized within schools with means equal to zero and standard deviations equal to one; the other dimensions of academic functioning were not included given space limitations, but information on them may be obtained by contacting the first author (see also the Results section).

to handling missing data under a wider range of conditions than does not only list- and pairwise deletion but also mean-imputation (Arbuckle, 1996; Wotheke, 2000). To identify differences between the United States and China, two-group nested model comparisons were employed: The unconstrained models were compared to more parsimonious models with constraints of equal coefficients imposed between the two countries on the effects of interest; for each set of models, the constraints were imposed one by one and then simultaneously. A significant difference ($\Delta\chi^2$) between an unconstrained model and a more parsimonious constrained model indicates a country difference. The same two parcels or items used in the CFAs conducted to establish measurement invariance (see the Method section) were employed for the latent constructs in the model; for grades, the four subjects were each used as indicators of the latent construct. A separate set of models was conducted for each dimension of academic functioning.

Prior research using this data set already established the total effects of parents' involvement on children's engagement and grades over time. C. S. Cheung and Pomerantz (2012) conducted sets of two-group nested SEM analyses examining if parents' involvement is predictive of children's academic functioning over the four waves (see also C. S. Cheung & Pomerantz, 2011): The effect of parents' involvement at Wave 1 on children's academic functioning (i.e., engagement and grades) at Wave 4 was evaluated, taking into account residual variance by adjusting for children's earlier (Wave 1) academic functioning as well as allowing the variance of parents' involvement and children's academic functioning at Wave 1 to correlate. The unconstrained, χ^2 s($df > 5$) < 3.91, CFI > .96, TLI > .95, RMSEA < .05, and constrained, χ^2 s($dfs = 4$) = 1.67, CFI > .96, TLI > .96, RMSEA = .04, models fit the data well, with the effects similar in the United States and China, $\Delta\chi^2$ s($df = 1$) < 1.5. The more involved parents were in children's learning, the more children were engaged ($\gamma = .15$ for self-regulated learning strategies and .08 for time spent on schoolwork; $ts > 2.66$, $ps < .01$), and the better their grades ($\gamma = .07$; $t = 3.01$, $p < .01$) 2 years later over and above their earlier engagement and grades.

In the current report, we used two-group nested SEM analyses to identify the role of the two value development pathways (i.e., the perception-acceptance and experience pathways) in explaining the effects of parents' involvement on children's academic functioning. As shown in Figure 2, children's reports of parents' involvement at Wave 1 were specified to predict children's perceptions of the value parents place on school achievement (i.e., the

first step of the perception-acceptance pathway, path a in Figures 1 and 2) and the value children themselves place on school (i.e., the first step of the experience pathway, path c in Figure 1 and c1 in Figure 2) at Wave 2. For the perception-acceptance pathway, children's perceptions of parents' values at Wave 2 were specified to predict their own values at Wave 3 (path b in Figures 1 and 2), which in turn were specified to predict children's academic functioning at Wave 4 (path d in Figures 1 and 2). For the experience pathway, the value children place on school achievement at Wave 2 was specified to predict the maintenance of such value at Wave 3 (path c2 in Figure 2), which in turn was specified to predict their academic functioning at Wave 4 (path d in Figures 1 and 2).

The mediating roles of the two pathways were simultaneously evaluated to assess the unique effects of each. Residual variance for each of the downstream constructs was taken into account. Specifically, as shown in Figure 2, corresponding constructs assessed 6 months prior to each of the constructs specified in the pathways were included to take into account autoregressive effects. Concurrent associations between constructs were also taken into account by allowing the variances (Wave 1) or error variances (Wave 2, 3, and 4) of the constructs to correlate within each wave. Because children's residence in single-headed household as well as their gender are associated with children's achievement during middle school (e.g., Downey, 1994; Dwyer & Johnson, 1997; Entwistle, 1997), children's reports of whether they reside with both parents in the same household (1 = *not residing with both parents*, 2 = *residing with both parents*) and their gender (1 = *boys*, 2 = *girls*) were included as covariates by specifying them to predict children's grades at Wave 4.

The unconstrained models (i.e., individual models for self-regulated learning strategies, time spent on schoolwork outside of school, and grades) fit the data adequately, χ^2 s($dfs > 319$) = 1080, CFI > .94, TLI > .92, RMSEA < .08. Two-group nested model comparisons indicated that the links comprising both the perception-acceptance and experience pathways were similar in the United States and China, $\Delta\chi^2$ s($dfs = 1$) < 2.2, ns; thus, all such effects were constrained to be equal between the two countries in the final constrained models, χ^2 s($dfs > 314$) > 1069, CFI > .96, TLI > .94, RMSEA < .07. As shown in Table 2 and Figure 2, there was support for the perception-acceptance pathway. Children's reports of parents' involvement at Wave 1 predicted children's perceptions of the value parents place on school achievement at Wave 2 taking into account children's earlier (Wave 1) perceptions ($ts = 2.96$, $ps < .01$). In turn, the more

Table 2
Summary of Model Fit and Parameter Estimates for the Value Development Models

Dimension of academic functioning	Model fit			Estimates			Delta coefficient	
	CFI	TLI	RMSEA	Path a (Involvement → Parent Value)	Path b (Parent Value → Child Value)	Path c (Child Value → Adjustment)	Perception-Acceptance Pathway	Experience Pathway
Grades	.95	.94	.07	.15**	.11**	.16***	2.08*	3.38**
SRL	.96	.94	.07	.24***	.17**	.19***	2.21*	5.48***
Time on schoolwork	.94	.94	.07	.23***	.19**	.10**	1.81†	4.55***

Note. CFI = comparative fit index; TLI = Tucker-Lewis index; RMSEA = root-mean-square error of approximation; SRL = self-regulated learning. Based on the chi-square difference tests, all paths comprising the indirect pathways were constrained to be equal between the United States and China. Estimates from the final constrained models are reported.

† $p < .06$. * $p < .05$. ** $p < .01$. *** $p < .001$.

children perceived parents as valuing school achievement at Wave 2, the more they themselves valued it at Wave 3 taking into account the earlier (Wave 2) value children placed on school achievement ($ts = 2.02, ps < .05$). The value children placed on school achievement at Wave 3 predicted enhanced engagement ($ts > 3.10, ps < .01$) and grades ($ts = 4.92, ps < .001$) among children at Wave 4 over and above their earlier (Wave 1) engagement and grades. Notably, at Wave 3, neither parents' involvement nor children's perceptions of the value parents place on school achievement uniquely predicted children's engagement or grades ($ts < 1$). Thus, although the value children placed on school achievement predicted their subsequent perceptions of the value parents place on it (see Figure 2), this was not a viable pathway by which parents' involvement benefits children's academic functioning.

There was also support for the experience pathway (see Figure 2 and Table 2). Parents' involvement as reported by children at Wave 1 predicted the value children themselves placed on school achievement at Wave 2 taking into account children's earlier (Wave 1) value ($ts = 2.08, ps < .05$). The value children placed on school achievement was maintained over time—that is from Wave 2 to 3 ($ts = 6.28, ps < .01$), which, as reported above, predicted children's engagement and grades at Wave 4.

The total effects of parents' involvement (Wave 1) on children's academic functioning (Wave 4) were no longer evident in either country ($\gamma_s < .03$) with the inclusion of the value development pathways, which resulted in a reduction of at least 65% of the total effect for each of the three dimensions of children's academic functioning. Mplus's delta method indicated that the two-step perception-acceptance pathway was significant in the United States and China in explaining the role of involvement in children's engagement as reflected in their self-regulated learning strategies ($zs > 2.21, ps < .05$) and grades ($zs > 2.08, ps < .05$). For engagement, as reflected in children's time spent on school-work outside of the school, the perception-acceptance pathway was marginal ($zs = 1.81, ps < .06$). The one-step experience pathway was evident across all three dimensions of children's academic functioning ($zs > 3.38, ps < .01$). These results are consistent with those yielded by analyses using bootstrap resampling techniques. For example, in the model focusing on grades as the final outcome, the estimate of the perception-acceptance pathway via perceptions of parental value and child value using 5,000 bootstrap resamples was .004 (95% CI = .001, .009), and that of the experience pathway was .017 (95% CI = .001, .042).

The model examined also allowed us to test the viability of alternative pathways—for example, the possibility that the value children place on school achievement leads them to report parents as more involved over time, which in turn leads children to view parents as more invested in their achievement, thereby enhancing children's academic functioning. To examine these alternative explanations, we evaluated the role of all possible pathways in the link between parents' involvement and children's academic functioning, including the value development pathways, simultaneously in the same model. The unconstrained, χ^2 's ($dfs > 360$) = 1203, CFI's > .92, TLI's > .90, RMSEA = .08, and constrained, χ^2 's ($dfs > 350$) > 1191, CFI's > .92, TLI's > .90, RMSEA < .08, models fit the data adequately. When simultaneously evaluated in the model with the two value development pathways, none of the alternative pathways (out of six possible pathways) was evident

($zs < 1.10, ns$). However, the two value development pathways remained significant ($zs > 2.10, ps < .05$), reflecting their uniqueness. Although none of the alternative pathways were evident, one link comprising one of them was: In both countries, children's value at Wave 2 was predictive of their perceptions of parents' value at Wave 3, adjusting for children's earlier perceptions ($\gamma_s = .25-.28, ts > 2.78, ps < .05$).

Discussion

The current research is the first empirical test of one of the most frequently proposed pathways—that is, value development—argued to underlie the benefits of parents' involvement in children's learning (e.g., Epstein, 1988; Grolnick & Slomiaczek, 1994). Consistent with the two-step value transmission model put forth by Grusec and Goodnow (1994), there was evidence for a *perception-acceptance pathway*: The more involved parents were—as reported by children—the more children perceived them as placing heightened value on school achievement; this, in turn, was predictive of children coming to value school achievement more over time. In line with ideas that parents' involvement may create experiences that foster value development among children (e.g., Hill & Tyson, 2009), there was also evidence that parents' involvement contributes directly to the value children place on school achievement (i.e., the *experience pathway*). Both pathways uniquely accounted for the beneficial effect of parents' involvement on children's later academic functioning (i.e., engagement and grades).

The effects of the two value development pathways were robust in that they remained even when alternative pathways were taken into account (e.g., the more children value school achievement, the more they see parents as valuing it, which heightens children's reports of parents' involvement, thereby enhancing their achievement); the value development pathways were also not due to children's gender or residence with both (vs. one) parent, which have been linked to children's achievement (e.g., Downey, 1994; Dwyer & Johnson, 1997; Entwistle, 1997). Although comparable to those of prior research using stringent statistical controls to identify indirect pathways over time (e.g., Davies, Woitach, Winter, & Cummings, 2008; NICHD Early Child Care Research Network, 2003), the effects of the value development pathway were modest—perhaps because value development has been underway for some time once children reach adolescence with only incremental change occurring at this time. Even modest effects, however, may be critical to offsetting the devaluing of school that often occurs among children over adolescence (for a review, see Wigfield & Wagner, 2005). Moreover, incremental change can be meaningful as it may accumulate over time (Pomerantz, Qin, Wang, & Chen, 2011). Moderation may also contribute to the modest effects. For example, drawing from Grusec and Goodnow (1994), when children have poor relationships with parents, parents' involvement in their learning may be less likely to lead them to take on parents' values as their own.

Increasingly research has focused on understanding the processes underlying the benefits of parents' involvement in children's learning for children's academic functioning (for a review, see Pomerantz et al., 2012). In this vein, children's actual and perceived competencies have been identified as important mechanisms (e.g., Dearing et al., 2006; Senechal & LeFevre, 2002).

Although it is possible that such mechanisms are distinct from the value development pathways identified in the current research, it is also possible that they work together. Value development may establish the foundation for growth in children's competencies: Once children come to see achievement in school as personally important, they may be more receptive to parents' instruction, which may develop their competencies, thereby allowing them to feel confident. Other mechanisms may also be a part of the value development pathways. For example, C. S. Cheung and Pomerantz (2012) found that the effect of parents' involvement on children's achievement was due in part to children adopting parent-oriented reasons (e.g., to meet parents' expectations) for school achievement; such motivation may be particularly likely to develop once children see parents as valuing school achievement, ultimately leading children to view achievement as personally important so that they may do well to satisfy parents who have committed substantial resources to their learning.

Despite differences in the quantity and quality of American and Chinese parents' involvement in children's learning (for reviews, see Chao & Tseng, 2002; Pomerantz, Ng, & Wang, 2008), the value development pathways were similarly evident in the United States and China. Although Chinese parents tend to accompany their involvement in children's learning with control more than do American parents (e.g., C. S. Cheung & Pomerantz, 2011), the more parents were involved, the more children viewed them as valuing school achievement and valued it themselves in both the United States and China. Moreover, children's perceptions of the value parents place on school achievement were similarly predictive over time of the value children themselves placed on it in the two countries. Thus, it appears that regardless of the quantity or quality, parents' involvement in children's learning may be a unique dimension of parenting in that it conveys parents' values while also having characteristics such as emotional support that may increase the accuracy of children's perceptions of such values as well as their acceptance of them.

The current research was guided by Grusec and Goodnow's (1994) two-step process model by which parents transmit their values to children. However, it diverged from the model in that the actual value parents place on school achievement was not directly assessed, but rather assumed to be reflected in parents' involvement in children's learning. Although parents' values likely drive their involvement, so do other forces—for example, children's invitations to be involved, parents' beliefs about their capacity to support children's learning, and whether parents see it as their role to be involved (for a review, see Hoover-Dempsey & Sandler, 1997). The current research did not examine the *accuracy* of value transmission, but rather what parents' involvement conveyed to children about the value parents' place on achievement in school. It is of note that the value parents place on children's school achievement may not be conveyed if parents are not involved. Hence, simply valuing school achievement may not reap the same benefits as being involved.

Limitations

Several limitations should be considered in interpreting the results. Perhaps most significantly, with the exception of grades, children served as the sole reporters. To rule out informant bias,

our model controlled for the concurrent associations between the child-reported constructs as well as the stability of each over time as both these links are likely to contain informant bias. Given such controls, the value development pathways are unlikely to contain informant bias. However, we went further in ruling out the possibility of other pathways that could result in bias due to children's reports—for example, we ensured that the effects were not simply due to children's values driving their reports of parents' values and involvement. Despite the merits of using multiple informants, it was crucial that children report on both their perceptions of the value parents place on school as well as the value they themselves place on school given that these constructs represent children's beliefs to which they likely have the best access. Yet, because children's reports of parents' involvement are only modestly associated with parents and teachers' reports (e.g., Bakker et al., 2007; Hill et al., 2004; Reynolds, 1992), it will be important for future research to examine the value development pathways using parents and teachers' reports.

The current research also did not distinguish between mothers and fathers' involvement, asking children instead to report on involvement as practiced by parents as a single entity. It is quite possible that mothers and fathers are differentially involved reflecting differences in their time and values. For example, Roest, Dubas, Gerris, and Engels (2009) reported only modest correspondence between Dutch mothers and fathers' values in terms of such things as the importance of pursuing happiness and working hard. Research in the United States indicates that mothers and fathers' involvement in children's learning does not necessarily overlap—for example, mothers are often more likely than fathers to attend school events and assist children with homework (Nord & West, 2001). Future research should examine if mothers and fathers' involvement differentially guides value development among children. Attention should also be given to the moderating role of the consistency between mothers and fathers in their values and involvement because when there is more agreement between parents in their values, children often have more accurate perceptions of parents' values (Knafo & Schwartz, 2004).

Given their homogeneity (e.g., the American sample was mainly of European descent and the Chinese sample was mainly of *Han* descent), the samples used in the current research do not represent the diversity of the United States and China. Thus, questions remain concerning within-culture variations in the role of value development in the effect of parents' involvement in children's learning. Within the United States, there is some evidence that how parents are involved varies demographically (e.g., Hill & Taylor, 2004; Snyder & Dillow, 2012). For example, the more educated parents are, the more they take part in events at children's school (Snyder & Dillow, 2012). It is possible that different types of involvement convey different messages about the value parents' place on school (e.g., those that children see as taking more time and energy indicate most that parents view school as important). Of additional concern, is that urban areas such as Beijing in China have been increasingly exposed to Western values in the past few decades. Thus, it is possible that the Chinese children in the current research interpret parents' involvement more similarly to American children than do Chinese children residing in rural areas.

Conclusions

Despite these limitations, the current research is of import in providing empirical support for a value development model of the effects of parents' involvement in children's learning. Such involvement appears to benefit children in part because it leads children to view school achievement as valuable, which heightens their engagement in school, ultimately enhancing their grades. Via a perception-acceptance pathway, when parents become involved in children's learning, children perceive parents as placing heightened value on achievement in school; such perceptions in turn foreshadow children viewing achievement in school as personally important. In an experience pathway, parents' involvement foreshadows children placing heightened value on school achievement presumably due to the experiences created by parents' involvement (e.g., discussion about school allows children to generate reasons for its utility), which in turn predicts enhanced academic functioning among children. These value development pathways were similarly evident in the United States and China where the quantity and quality of parents' involvement differ.

References

Alexander, K. L., Entwistle, D. R., & Dauber, S. L. (1993). First-grade classroom behavior: Its short- and long-term consequences for school performance. *Child Development*, 64, 801–814. doi:10.2307/1131219

Arbuckle, J. L. (1996). Full information estimation in the presence of incomplete data. In G. A. Marcoulides & R. E. Schumacker (Eds.), *Advanced structural equation modeling: Issues and techniques* (pp. 243–277). Mahwah, NJ: Erlbaum.

Arnett, J. J. (2008). The neglected 95%: Why American psychology needs to become less American. *American Psychologist*, 63, 602–614. doi:10.1037/0003-066X.63.7.602

Bakker, J., Denessen, E., & Brus-Laeven, M. (2007). Socio-economic background, parental involvement and teacher perceptions of these in relation to pupil achievement. *Educational Studies*, 33, 177–192. doi:10.1080/03055690601068345

Barni, D., Ranieri, S., Scabini, E., & Rosnati, R. (2011). Value transmission in the family: Do adolescents accept the values their parents want to transmit? *Journal of Moral Education*, 40, 105–121. doi:10.1080/03057240.2011.553797

Beijing Municipal Bureau of Statistics. (2005). *Beijing statistical yearbook 2005*. Retrieved from <http://www.bjstats.gov.cn/tjnj/2005-tjnj/>

Bem, D. J. (1967). Self-perception: An alternative interpretation of cognitive dissonance phenomena. *Psychological Review*, 74, 183–200. doi:10.1037/h0024835

Bem, D. J. (1972). Self-perception theory. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 6, pp. 1–62). New York, NY: Academic Press.

Brislin, R. W. (1980). Translation and content analysis of oral and written materials. In H. C. Triandis & J. W. Berry (Eds.), *Handbook of cross-cultural psychology: Vol. 2. Methodology* (pp. 389–444). Boston, MA: Allyn & Bacon.

Chao, R. K. (2000). The parenting of immigrant Chinese and European American mothers: Relations between parenting styles, socialization goals, and parental practices. *Journal of Applied Developmental Psychology*, 21, 233–248. doi:10.1016/S0193-3973(99)00037-4

Chao, R., & Tseng, V. (2002). Parenting of Asians. In M. H. Bornstein (Ed.), *Handbook of parenting: Vol. 4 Social conditions and applied parenting* (2nd ed., pp. 59–93). Mahwah, NJ: Erlbaum.

Chen, C. S., & Stevenson, H. W. (1989). Homework: A cross-cultural examination. *Child Development*, 60, 551–561. doi:10.2307/1130721

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling*, 14, 464–504. doi:10.1080/10705510701301834

Cheung, C. S., & Pomerantz, E. M. (2011). Parents' involvement in the United States and China: Implications for children's academic and emotional adjustment. *Child Development*, 82, 932–950. doi:10.1111/j.1467-8624.2011.01582.x

Cheung, C. S., & Pomerantz, E. M. (2012). Why does parents' involvement enhance children's achievement? The role of parent-oriented motivation. *Journal of Educational Psychology*, 104, 820–832. doi:10.1037/a0027183

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling*, 9, 233–255. doi:10.1207/S15328007SEM0902_5

China Population and Development Research Center. (2001). *Major figures of the 2000 census*. Retrieved from <http://www.cprc.org.cn/en/e5cndata1.htm>

Coe, D. P., Pivarnik, J. M., Womack, C. J., Reeves, M. J., & Malina, R. M. (2006). Effect of physical education and activity levels on academic achievement in children. *Medicine and Science in Sports and Exercise*, 38, 1515–1519. doi:10.1249/01.mss.0000227537.13175.1b

Davies, P. T., Woitach, M. J., Winter, M. A., & Cummings, E. M. (2008). Children's insecure representations of the interparental relationship and their school adjustment: The mediating role of attention difficulties. *Child Development*, 79, 1570–1582. doi:10.1111/j.1467-8624.2008.01206.x

Dearing, E., Kreider, H., Simpkins, S., & Weiss, H. B. (2006). Family involvement in school and low-income children's literacy performance: Longitudinal associations between and within families. *Journal of Educational Psychology*, 98, 653–664. doi:10.1037/0022-0663.98.4.653

Deslandes, R., Bouchard, P., & St.-Amant, J. (1998). Family variables as predictors of school achievement: Sex differences in Quebec adolescents. *Canadian Journal of Education*, 23, 390–404. doi:10.2307/1585754

Downey, D. (1994). The school performance of children from single-mother and single-father families: Economic or interpersonal deprivation? *Journal of Family Issues*, 15, 129–147. doi:10.1177/019251394015001006

Dowson, M., & McInerney, D. M. (2004). The development and validation of the Goal Orientation and Learning Strategies Survey (GOALS-S). *Educational and Psychological Measurement*, 64, 290–310. doi:10.1177/0013164403251335

Duckworth, A. L., & Seligman, M. E. P. (2005). Self-discipline outdoes IQ in predicting academic performance of adolescents. *Psychological Science*, 16, 939–944. doi:10.1111/j.1467-9280.2005.01641.x

Dwyer, C., & Johnson, L. (1997). Grades, accomplishments, and correlates. In W. Willingham & N. Cole (Eds.), *Gender and fair assessment* (pp. 127–156). Mahwah, NJ: Erlbaum.

Eccles, J. S., Adler, T. F., Futterman, R., Goff, S. B., Kaczala, C. M., Meece, J. L., & Midgley, C. (1983). Expectancies, values, and academic behaviors. In J. T. Spence (Ed.), *Achievement and achievement motives: Psychological and sociological approaches* (pp. 75–146). San Francisco, CA: Freeman.

Entwistle, D. R. (1997). *Children, schools, and inequality*. Boulder, CO: Westview Press.

Epstein, J. L. (1988). How do we improve programs for parental involvement? *Educational Horizons*, 66, 75–77.

Erkut, S. (2010). Developing multiple language versions of instruments for intercultural research. *Child Development Perspectives*, 4, 19–24. doi:10.1111/j.1750-8606.2009.00111.x

Fuligni, A. J., Tseng, V., & Lam, M. (1999). Attitudes toward family obligation among American adolescents with Asian, Latin American, and European American backgrounds. *Child Development*, 70, 1030–1044. doi:10.1111/1467-8624.00075

Geiser, S., & Santelices, M. V. (2007). *Validity of high-school grades in predicting student success beyond the freshman year: High-school record vs standardized tests as indicators of four-year college outcomes*. Berkeley, CA: Center for Studies in Higher Education, University of California, Berkeley.

Grolnick, W. S., Friendly, R., & Bellas, V. (2009). Parenting and children's motivation at school. In K. Wentzel & A. Wigfield (Eds.), *Handbook of motivation at school* (pp. 279–300). Mahwah, NJ: Erlbaum.

Grolnick, W. S., Ryan, R. M., & Deci, E. L. (1997). Internalization in the family: The self-determination perspective. In J. E. Grusec & L. Kuczynski (Eds.), *Parenting and children's internalization of values* (pp. 135–161). New York, NY: Wiley.

Grolnick, W. S., & Slowiaczek, M. L. (1994). Parents' involvement in children's schooling: A multidimensional conceptualization and motivational model. *Child Development*, 65, 237–252. doi:10.2307/1131378

Grusec, J. E. (2002). Parenting socialization and children's acquisition of values. In M. H. Bornstein (Ed.), *Handbook of parenting: Vol. 5: Practical issues in parenting* (pp. 143–167). Mahwah, NJ: Erlbaum.

Grusec, J. E., & Goodnow, J. J. (1994). Impact of parental discipline methods on the child's internalization of values: A reconceptualization of current points of view. *Developmental Psychology*, 30, 4–19. doi:10.1037/0012-1649.30.1.4

Henrich, J., Heine, S. J., & Norenzayan, A. (2010a). Most people are not WEIRD. *Nature*, 466, 29. doi:10.1038/466029a

Henrich, J., Heine, S. J., & Norenzayan, A. (2010b). The weirdest people in the world. *Behavioral and Brain Sciences*, 33, 61–83. doi:10.1017/S0140525X0999152X

Hill, N. E., Castellino, D. R., Lansford, J. E., Nowlin, P., Dodge, K. A., Bates, J. E., & Petit, G. S. (2004). Parent-academic involvement as related to school behavior, achievement, and aspirations: Demographic variations across adolescence. *Child Development*, 75, 1491–1509. doi:10.1111/j.1467-8624.2004.00753.x

Hill, N. E., & Taylor, L. C. (2004). Parental school involvement and children's academic achievement: Pragmatics and issues. *Current Directions in Psychological Science*, 13, 161–164. doi:10.1111/j.0963-7214.2004.00298.x

Hill, N. E., & Tyson, D. (2009). Parental involvement in middle school: A meta-analytic assessment of the strategies that promote achievement. *Developmental Psychology*, 45, 740–763. doi:10.1037/a0015362

Hoover-Dempsey, K. V., & Sandler, H. (1997). Why do parents become involved in their children's education? *Review of Educational Research*, 67, 3–42. doi:10.3102/00346543067001003

Jeynes, W. H. (2005). A meta-analysis of the relation of parental involvement to urban elementary school student academic achievement. *Urban Education*, 40, 237–269. doi:10.1177/0042085905274540

Jeynes, W. H. (2007). The relationship between parental involvement and urban secondary school student academic achievement: A meta-analysis. *Urban Education*, 42, 82–110. doi:10.1177/0042085906293818

Keith, T. Z. (2006). *Multiple regression and beyond*. Boston, MA: Allyn & Bacon.

Kenney-Benson, G., Pomerantz, E. M., Ryan, A., & Patrick, H. (2006). Sex differences in math performance: The role of how children approach school. *Developmental Psychology*, 42, 11–26. doi:10.1037/0012-1649.42.1.11

Kerr, M., & Stattin, H. (2000). What parents know, how they know it, and several forms of adolescent adjustment: Further support for a reinterpretation of monitoring. *Developmental Psychology*, 36, 366–380. doi:10.1037/0012-1649.36.3.366

Kline, R. B. (1998). *Principles and practice of structural equation modeling*. New York, NY: Guilford Press.

Knafo, A., & Schwartz, S. H. (2004). Identity formation and parent-child value congruence in adolescence. *British Journal of Developmental Psychology*, 22, 439–458. doi:10.1348/0261510041552765

Knafo, A., & Schwartz, S. H. (2009). Accounting for parent-child value congruence: Theoretical considerations and empirical evidence. In U. Schönflug (Ed.), *Culture and psychology* (pp. 240–268). New York, NY: Cambridge University Press. doi:10.1017/CBO9780511804670.012

Kohl, G. O., Lengua, L. J., McMahon, R. J., & The Conduct Problems Prevention Research Group. (2000). Parent involvement in school: Conceptualizing multiple dimensions and their relations with family and demographic risk factors. *Journal of School Psychology*, 38, 501–523. doi:10.1016/S0022-4405(00)00050-9

Little, T. D. (1997). Mean and covariance structures (MACS) analyses of cross-cultural data: Practical and theoretical issues. *Multivariate Behavioral Research*, 32, 53–76. doi:10.1207/s15327906mbr3201_3

Little, T. D., Cunningham, W. A., Shahar, G., & Widaman, K. F. (2002). To parcel or not to parcel: Exploring the question, weighing the merits. *Structural Equation Modeling*, 9, 151–173. doi:10.1207/S15328007SEM0902_1

Little, T. D., Rhemtulla, M., Gibson, K., & Schoemann, A. M. (2013). Why the items versus parcels controversy needn't be one. *Psychological Methods*, 18, 285–300. doi:10.1037/a0033266

McDonald, R. P., & Ho, M.-H. R. (2002). Principles and practice in reporting structural equation analyses. *Psychological Methods*, 7, 64–82. doi:10.1037/1082-989X.7.1.64

Muthén, L. K., & Muthén, B. O. (1998–2012). *Mplus user's guide* (7th ed.). Los Angeles, CA: Muthén & Muthén.

Ng, F. F., Pomerantz, E. M., & Lam, S. F. (2007). European American and Chinese parents' responses to children's success and failure: Implications for children's responses. *Developmental Psychology*, 43, 1239–1255. doi:10.1037/0012-1649.43.5.1239

NICHD Early Child Care Research Network. (2003). Do children's attention processes mediate the link between family predictors and school readiness? *Developmental Psychology*, 39, 581–593. doi:10.1037/0012-1649.39.3.581

Nord, C. W., & West, J. (2001). *Fathers and mothers' involvement in their children's schools by family type and resident status*. Washington, DC: National Center for Educational Statistics.

Okagaki, L., & Bevis, C. (1999). Transmission of religious values: Relations between parents' and daughters' beliefs. *The Journal of Genetic Psychology: Research and Theory on Human Development*, 160, 303–318. doi:10.1080/00221329909595401

Pintrich, P. R. (1999). The role of motivation in promoting and sustaining self-regulated learning. *International Journal of Educational Research*, 31, 459–470. doi:10.1016/S0883-0355(99)00015-4

Pomerantz, E. M., Kim, E. M., & Cheung, C. S. (2012). Parents' involvement in children's learning. In K. R. Harris, S. Graham, T. C. Urdan, S. Graham, J. M. Royer, & M. Zeidner (Eds.), *APA educational psychology handbook* (pp. 417–440). Washington, DC: American Psychological Association. doi:10.1037/13274-017

Pomerantz, E. M., Ng, F. F., Cheung, C. S., & Qu, Y. (in press). How to raise happy children who succeed in school: Lessons from China and the United States. *Child Development Perspectives*. doi:10.1111/cdep.12063

Pomerantz, E. M., Ng, F. F., & Wang, Q. (2008). Culture, parenting, and motivation: The case of East Asia and the United States. In M. L. Maehr, S. A. Karabenick, & T. C. Urdan (Eds.), *Advances in motivation and achievement: Social psychological perspectives* (Vol. 15, pp. 209–240). Bingley, England: Emerald Group. doi:10.1016/S0749-7423(08)15007-5

Pomerantz, E. M., Qin, L., Wang, Q., & Chen, H. (2009). American and Chinese early adolescents' inclusion of their relationships with their parents in their self-construals. *Child Development*, 80, 792–807. doi:10.1111/j.1467-8624.2009.01298.x

Pomerantz, E. M., Qin, L., Wang, Q., & Chen, H. (2011). Changes in early adolescents' sense of responsibility to their parents in the United States and China: Implications for their academic functioning. *Child Development*, 82, 1136–1151. doi:10.1111/j.1467-8624.2011.01588.x

Pomerantz, E. M., Saxon, J. L., & Oishi, S. (2000). The psychological trade-offs of goal investment. *Journal of Personality and Social Psychology*, 79, 617–630. doi:10.1037/0022-3514.79.4.617

Reynolds, A. J. (1992). Comparing measures of parental involvement and their effects on academic achievement. *Early Childhood Research Quarterly*, 7, 441–462. doi:10.1016/0885-2006(92)90031-S

Rhemtulla, M., Brosseau-Liard, P., & Savalei, V. (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under sub-optimal conditions. *Psychological Methods*, 17, 354–373. doi:10.1037/a0029315

Roest, A. M. C., Dubas, J. S., Gerris, J. R. M., & Engels, R. C. M. E. (2009). Value similarities among fathers, mothers, and adolescents and the role of a cultural stereotype: Different measurement strategies reconsidered. *Journal of Research on Adolescence*, 19, 812–833. doi:10.1111/j.1532-7795.2009.00621.x

Schwartz, D., Kelly, B. M., & Duong, M. T. (2013). Do academically-engaged adolescents experience social sanctions from the peer group? *Journal of Youth and Adolescence*, 42, 1319–1330. doi:10.1007/s10964-012-9882-4

Sénéchal, M., & LeFevre, J. (2002). Parental involvement in the development of children's reading skill: A five year longitudinal study. *Child Development*, 73, 445–460. doi:10.1111/1467-8624.00417

Snyder, T. D., & Dillow, S. A. (2012). *Digest of education statistics 2011* (NCES 2012-001). Washington, DC: National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education.

Stattin, H., & Kerr, M. (2000). Parental monitoring: A reinterpretation. *Child Development*, 71, 1072–1085. doi:10.1111/1467-8624.00210

U.S. Census Bureau. (2007). *Census 2000 Summary File 1 and 3*. Retrieved January 4, 2010, from http://factfinder.census.gov/servlet/DatasetMainPageServlet?_lang=en&_ts=235058466046&_ds_name=DEC_2000_SF3_U&_program=

Wang, M.-T., & Fredricks, J. (2014). The reciprocal links between school engagement and youth problem behavior during adolescence. *Child Development*, 85, 722–737. doi:10.1111/cdev.12138

Wang, M.-T., & Peck, S. (2013). Adolescent educational success and mental health vary across school engagement profiles. *Developmental Psychology*, 49, 1266–1276. doi:10.1037/a0030028

Wang, Q., & Pomerantz, E. M. (2009). The motivational landscape of early adolescence in the United States and China: A longitudinal study. *Child Development*, 80, 1272–1287. doi:10.1111/j.1467-8624.2009.01331.x

Wigfield, A., & Wagner, A. L. (2005). Competence, motivation, and identity development during adolescence. In A. Elliott & C. Dweck (Eds.), *Handbook of competence and motivation* (pp. 222–239). New York, NY: Guilford Press.

Wood, R. E., & Locke, E. A. (1987). The relation of self-efficacy and grade goals to academic performance. *Educational and Psychological Measurement*, 47, 1013–1024. doi:10.1177/0013164487474017

Wothke, W. (2000). Longitudinal and multigroup modeling with missing data. In T. D. Little, K. U. Schnabel, & J. Baumert (Eds.), *Modeling longitudinal and multilevel data: Practical issues, applied approaches, and specific examples* (pp. 219–240). Mahwah, NJ: Erlbaum.

Received March 14, 2013

Revision received June 11, 2014

Accepted June 17, 2014 ■

E-Mail Notification of Your Latest Issue Online!

Would you like to know when the next issue of your favorite APA journal will be available online? This service is now available to you. Sign up at <http://notify.apa.org/> and you will be notified by e-mail when issues of interest to you become available!

Instructions to Authors
Journal of Educational Psychology
www.apa.org/pubs/journals/edu

The main purpose of the *Journal of Educational Psychology* is to publish original, primary psychological research pertaining to education across all ages and educational levels. A secondary purpose of the *Journal* is the occasional publication of exceptionally important theoretical and review articles that are pertinent to educational psychology.

Manuscript preparation. Authors should prepare manuscripts according to the *Publication Manual of the American Psychological Association* (6th ed.). Manuscripts may be copyedited for bias-free language (see pp. 70–77 of the *Publication Manual*). Formatting instructions (all copy must be double-spaced) and instructions on the preparation of tables, figures, references, metrics, and abstracts appear in the *Manual*. For APA's Checklist for Manuscript Submission, see www.apa.org/pubs/journals/edu. **Abstract and keywords.** All manuscripts must include an abstract containing a maximum of 250 words typed on a separate page. After the abstract, please supply up to five keywords or brief phrases. **References.** References should be listed in alphabetical order. Each listed reference should be cited in text, and each text citation should be listed in the References. Basic formats are as follows:

Hughes, G., Desantis, A., & Waszak, F. (2013). Mechanisms of intentional binding and sensory attenuation: The role of temporal prediction, temporal control, identity prediction, and motor prediction. *Psychological Bulletin*, 139, 133–151. <http://dx.doi.org/10.1037/a0028566>

Rogers, T. T., & McClelland, J. L. (2004). *Semantic cognition: A parallel distributed processing approach*. Cambridge, MA: MIT Press.

Gill, M. J., & Sypher, B. D. (2009). Workplace incivility and organizational trust. In P. Lutgen-Sandvik & B. D. Sypher (Eds.), *Destructive organizational communication: Processes, consequences, and constructive ways of organizing* (pp. 53–73). New York, NY: Taylor & Francis.

Adequate description of participants is critical to the science and practice of educational psychology; this allows readers to assess the results, determine generalizability of findings, and make comparisons in replications, extensions, literature reviews, or secondary data analyses. Authors should see guidelines for sample–subject description in the *Manual*. Appropriate indexes of effect size or strength of relationship should be incorporated in the results section of the manuscript (see p. 34 of the *Manual*). Information that allows the reader to assess not only the significance but also the magnitude of the observed effects or relationships clarifies the importance of the findings. **Figures.** Graphics files are welcome if supplied in TIFF or EPS format. APA's policy on publication of color figures is available at <http://www.apa.org/pubs/authors/instructions.aspx?item=6>.

Publication policies. APA policy prohibits an author from submitting the same manuscript for concurrent consideration by two or more publications. APA policy regarding posting articles on the Internet may be found at www.apa.org/pubs/authors/posting.aspx. In addition, it is a violation of APA Ethical Principles to publish "as original data, data that have been previously published" (Standard 8.13). As this is a primary journal that publishes original material only, APA policy prohibits publication of any manuscript or data that have already been published in

whole or substantial part elsewhere. Authors have an obligation to consult journal editors concerning prior publication of any data on which their article depends. In addition, APA Ethical Principles specify that "after research results are published, psychologists do not withhold the data on which their conclusions are based from other competent professionals who seek to verify the substantive claims through reanalysis and who intend to use such data only for that purpose, provided that the confidentiality of the participants can be protected and unless legal rights concerning proprietary data preclude their release" (Standard 8.14). Authors must have available their data throughout the editorial review process and for at least 5 years after the date of publication.

Masked review policy. The *Journal* has a masked review policy, which means that the identities of both authors and reviewers are masked. Every effort should be made by the authors to see that the manuscript itself contains no clues to their identities. Authors should never use first person (*I, my, we, our*) when referring to a study conducted by the author(s) or when doing so reveals the authors' identities, e.g., "in our previous work, Johnson et al., 1998 reported that . . ." Instead, references to the authors' work should be in third person, e.g., "Johnson et al. (1998) reported that . . ." The authors' institutional affiliations should also be masked in the manuscript. Authors submitting manuscripts are required to include in the cover letter the title of the manuscript along with all authors' names and institutional affiliations. However, the first page of the manuscript should omit the authors' names and affiliations, but should include the title of the manuscript and the date it is submitted. Responsibility for masking the manuscript rests with the authors; manuscripts will be returned to the author if not appropriately masked. If the manuscript is accepted, authors will be asked to make changes in wording so that the paper is no longer masked. Authors are required to state in writing that they have complied with APA ethical standards in the treatment of their sample, or to describe the details of treatment. A copy of the APA Ethical Principles may be obtained at www.apa.org/ethics/ or by writing the APA Ethics Office, 750 First Street, NE, Washington, DC 20002-4242. APA requires authors to reveal any possible conflict of interest in the conduct and reporting of research (e.g., financial interests in a test procedure, funding by pharmaceutical companies for drug research). Authors of accepted manuscripts will be required to transfer copyright to APA.

Permissions. Authors of accepted papers must obtain and provide to the editor on final acceptance all necessary permissions to reproduce in print and electronic form any copyrighted work, including test materials (or portions thereof), photographs, and other graphic images (including those used as stimuli in experiments). On advice of counsel, APA may decline to publish any image whose copyright status is unknown.

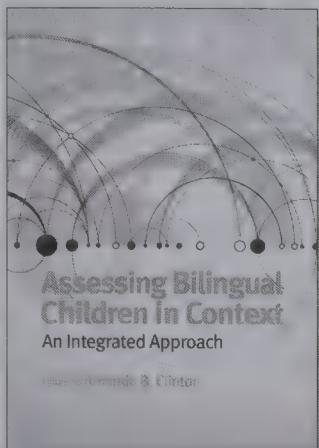
Supplemental materials. APA can place supplementary materials online, which will be available via the published article in the PsycARTICLES database. To submit such materials, please see www.apa.org/pubs/authors/supp-material.aspx for details. Authors of accepted papers will be asked to work with the editor and production staff to provide supplementary materials as appropriate.

Submission. Authors should submit their manuscripts electronically via the Manuscript Submission Portal at www.apa.org/pubs/journals/edu/index.aspx (follow the link for submission under Instructions to Authors). General correspondence may be addressed to the incoming editorial office at AConley@apa.org.

ASSESSING BILINGUAL CHILDREN IN CONTEXT

An Integrated Approach

Edited by Amanda B. Clinton



Children who are learning a second language and are referred for psychological assessment frequently present with unique personal backgrounds. They may have relocated from a Syrian refugee camp, immigrated from Mexico to escape poverty, or grown up navigating two languages spoken by biracial parents. Their individual histories can have long-term effects on learning, behavior, and social-emotional development.

This book was written to inform clinicians and educators working with youth who demonstrate variability in knowledge and fluency in either one or both of their two languages. It explores the influence of experiences like poverty and immigration on the biological processes of second-language acquisition, looks at the effects of cross-language transfer, discusses international and multicultural complexities critical to understanding the bilingual child, and examines the biological and neurological bases of second-language acquisition. The authors expertly synthesize this material, offering a set of guidelines for assessment. In addition to case studies that illustrate the application of the principles discussed, they provide concise graphic tools, such as checklists and charts, to provide readers with a succinct point of reference.

This is not a book about tests. Rather, it is a book about children and the complexity of evaluating their functioning when they are acquiring a second language, written for the professionals who wish to help by gaining a complete understanding of the contexts that shape them. **Series: Division 16 / School Psychology. 2014. 281 pages. Hardcover.**

List: \$69.95 | APA Member/Affiliate: \$49.95 | ISBN 978-1-4338-1556-2 | Item # 4317323

CONTENTS:

Introduction | 1. Challenges and Complexities in the Assessment of the Bilingual Student | I. Insights From Neuroscience and Cognitive Psychology on Cross-Language Transfer | 2. Cross-Language Transfer in Bilingual Students | 3. Neuropsychological Considerations in Bilingual Assessment: The Underlying Basis of Language Disability | 4. Implications of Semilingualism for Assessment of the Bilingual Child | II. Practical Implications for Assessment | 5. Integrated Intellectual Assessment of the Bilingual Student | 6. Response to Intervention and Bilingual Learners: Promises and Potential Problems | 7. Integrated Social-Emotional Assessment of the Bilingual Child | III. A New Vision: Integrating Concepts in Bilingual Assessment | 8. Acculturation and Sociocognitive Factors | 9. Assessing Bilingual Students' Writing | 10. Implications of Bilingualism in Reading Assessment | 11. An Integrated Approach to the Assessment of the Refugee Student

ALSO OF INTEREST

AVAILABLE ON AMAZON KINDLE®



The Immigrant Paradox in Children and Adolescents: Is Becoming American a Developmental Risk?
Edited by Cynthia García Coll and Amy Kerivan Marks
2012. 328 pages. Hardcover.

List: \$49.95 | APA Member/Affiliate: \$39.95
ISBN 978-1-4338-1053-4 | Item # 4318097

AVAILABLE ON AMAZON KINDLE®



Neuropsychological Assessment and Intervention for Youth: An Evidence-Based Approach to Emotional and Behavioral Disorders
Edited by Linda A. Reddy, Adam S. Weissman, and James B. Hale
2013. 364 pages. Hardcover.

List: \$69.95 | APA Member/Affiliate: \$49.95
ISBN 978-1-4338-1266-8 | Item # 4316149



Bilingualism and Cognition: Informing Research, Pedagogy, and Policy
Eugene E. García and José E. Náñez, Sr.
2011. 242 pages. Hardcover.

List: \$39.95 | APA Member/Affiliate: \$34.95
ISBN 978-1-4338-0879-1 | Item # 4318087



AMERICAN PSYCHOLOGICAL ASSOCIATION

APA BOOKS ORDERING INFORMATION: 800-374-2721 • www.apa.org/pubs/books

In Washington, DC, call: 202-336-5510 • TDD/TTY: 202-336-6123 • Fax: 202-336-5502
In Europe, Africa, or the Middle East, call: +44 (0) 1767 604972

AD2709

NEW RELEASES

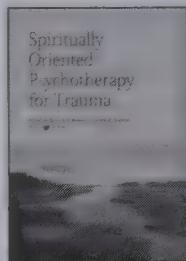
from the American Psychological Association



A Practical Guide to PTSD Treatment

Pharmacological and Psychotherapeutic Approaches
Edited by Nancy C. Bernardy and Matthew J. Friedman
2015. 192 pages. Paperback.

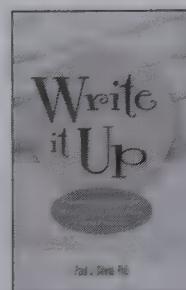
List: \$29.95 | APA Member/Affiliate: \$29.95
ISBN 978-1-4338-1832-5 | Item # 4317356



Spiritually Oriented Psychotherapy for Trauma

Edited by Donald E. Walker, Christine A. Courtois, and Jamie D. Aten
2015. 312 pages. Hardcover.

List: \$59.95 | APA Member/Affiliate: \$49.95
ISBN 978-1-4338-1816-5 | Item # 4317354



AN APA LIFETOOLS® BOOK Write It Up

Practical Strategies for Writing and Publishing Journal Articles
Paul Silvia
2015. 224 pages. Paperback.

List: \$19.95 | APA Member/Affiliate: \$19.95
ISBN 978-1-4338-1814-1 | Item # 4441024



Interdisciplinary Frameworks for Schools

Best Professional Practices for Serving the Needs of All Students
Virginia Wise Berninger
2015. 432 pages. Hardcover.

List: \$79.95 | APA Member/Affiliate: \$49.95
ISBN 978-1-4338-1808-0 | Item # 4317352

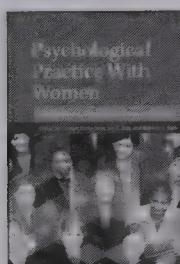


APA Handbook of Career Intervention

Volume 1: Foundations
Volume 2: Applications
Editors-in-Chief Paul J. Hartung, Mark L. Savickas, and W. Bruce Walsh
2015. 1,008 pages. Hardcover.

■ Series: APA Handbooks in Psychology®

List: \$395.00 | APA Member/Affiliate: \$195.00
ISBN 978-1-4338-1753-3 | Item # 4311514



Psychological Practice With Women

Guidelines, Diversity, Empowerment
Edited by Carolyn Zerbe Enns, Joy K. Rice, and Roberta L. Nutt
2015. 304 pages. Hardcover.

■ Series: Division 35: Psychology of Women

List: \$69.95 | APA Member/Affiliate: \$49.95
ISBN 978-1-4338-1812-7 | Item # 4317353



Premature Termination in Psychotherapy

Strategies for Engaging Clients and Improving Outcomes
Joshua K. Swift and Roger P. Greenberg
2015. 216 pages. Hardcover.

List: \$69.95 | APA Member/Affiliate: \$49.95
ISBN 978-1-4338-1801-1 | Item # 4317349

AMERICAN PSYCHOLOGICAL ASSOCIATION

TO ORDER: 800-374-2721 • www.apa.org/pubs/books

APA.org

SINGLE-CASE INTERVENTION RESEARCH

Methodological and Statistical Advances

Edited by Thomas R. Kratochwill and Joel R. Levin



Single Case Design (SCD) is a highly flexible method of conducting applied research where there is no control group/condition or possibility of collecting data from large groups of participants. Thanks to remarkable methodological and statistical advances in recent years, single case design (SCD) research has become a viable and often essential option for researchers in applied psychology, education, and related fields. This book, with contributions from leading experts, not only summarizes the state of the field today but offers the latest information and tools for researchers. It is a compendium of information and tools for researchers considering SCD research, a methodology in which a subject serves as the experimental control. **Series: Division 16: School Psychology. 2014. 408 pages.** *Hardcover.*

List: \$79.95 | APA Member/Affiliate: \$49.95 | ISBN 978-1-4338-1751-9 | Item # 4316163

CONTENTS

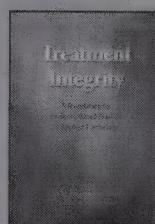
Introduction: An Overview of Single-Case Intervention Research, Thomas R. Kratochwill and Joel R. Levin | **I. Methodologies and Analyses** | Chapter 1. Constructing Single-Case Research Designs: Logic and Options, Robert H. Horner and Samuel L. Odom | Chapter 2. Enhancing the Scientific Credibility of Single-Case Intervention Research: Randomization to the Rescue, Thomas R. Kratochwill and Joel R. Levin | Chapter 3. Visual Analysis of Single-Case Intervention Research: Conceptual and Methodological Issues, Thomas R. Kratochwill, Joel R. Levin, Robert H. Horner, and Christopher M. Swoboda | Chapter 4. Non-Overlap Analysis for Single-Case Research, Richard I. Parker, Kimberly J. Vannest, and John L. Davis | Chapter 5. Single-Case Permutation and Randomization Statistical Tests: Present Status, Promising New Developments, John M. Ferron and Joel R. Levin | Chapter 6. The Single-Case Data-Analysis ExPRT (Excel® Package of Randomization Tests), Joel R. Levin, Anya S. Evmenova, and Boris S. Gafurov | Chapter 7. Using Multilevel Models to Analyze Single-Case Design Data, David M. Rindskopf and John M. Ferron | Chapter 8. Analyzing Single-Case Designs: d, G, Hierarchical Models, Bayesian Estimators, Generalized Additive Models, and the Hopes and Fears of Researchers about Analyses, William R. Shadish, Larry V. Hedges, James Pustejovsky, David M. Rindskopf, Jonathan G. Boyajian, and Kristynn J. Sullivan | Chapter 9. The Role of Single-Case Designs in Supporting Rigorous Intervention Development and Evaluation at the Institute of Education Sciences, Jacquelyn A. Buckley, Deborah L. Speece, and Joan E. McLaughlin | **II. Reactions from Leaders in the Field** | Chapter 10. Single-Case Designs and Large N Studies: The Best of Both Worlds, Susan M. Sheridan | Chapter 11. Using Single-Case Research Designs in Programs of Research, Ann P. Kaiser | Chapter 12. Reactions From Journal Editors: Journal of School Psychology, Randy G. Floyd | Chapter 13. One Editorial Perspective: School Psychology Quarterly, Randy W. Kamphaus | Chapter 14. Reflections from Journal Editors: School Psychology Review, Matthew K. Burns

ALSO OF INTEREST



APA Dictionary of Statistics and Research Methods
Editor-in-Chief Sheldon Zedeck
2014. 452 pages.
Hardcover.

List: \$39.95 | APA Member/Affiliate: \$29.95
ISBN 978-1-4338-1533-1 | Item # 4311019



Treatment Integrity: A Foundation for Evidence-Based Practice in Applied Psychology
Edited by Lisa M. Hagermoser Sanetti and Thomas R. Kratochwill
2014. 320 pages. *Hardcover.*

• Series: Division 16: School Psychology

List: \$69.95 | APA Member/Affiliate: \$49.95
ISBN 978-1-4338-1581-2 | Item # 4317327



AVAILABLE ON AMAZON KINDLE® Advanced Methods for Conducting Online Behavioral Research
Edited by Samuel D. Gosling and John A. Johnson
2010. 286 pages. *Hardcover.*

List: \$49.95 | APA Member/Affiliate: \$39.95
ISBN 978-1-4338-0695-7 | Item # 4311014



AMERICAN PSYCHOLOGICAL ASSOCIATION

APA BOOKS ORDERING INFORMATION: 800-374-2721 • www.apa.org/pubs/books

In Washington, DC, call: 202-336-5510 • TDD/TTY: 202-336-6123 • Fax: 202-336-5502
In Europe, Africa, or the Middle East, call: +44 (0) 1767 604972

AD2733

BEST SELLERS

from the American Psychological Association



AMERICAN
PSYCHOLOGICAL
ASSOCIATION

APA Handbook of Personality and Social Psychology

Volume 1: Attitudes and Social Cognition
Volume 2: Group Processes
Volume 3: Interpersonal Relations
Volume 4: Personality Processes and Individual Differences
Editors-in-Chief Mario Mikulincer and Phillip R. Shaver

2015. 3,056 pages. Hardcover.

Series: APA Handbooks in Psychology®

List: \$895.00 | APA Member/Affiliate: \$495.00
ISBN 978-1-4338-1699-4 | Item # 4311513

APA Handbook of Forensic Psychology

Volume 1: Individual and Situational Influences in Criminal and Civil Contexts
Volume 2: Criminal Investigation, Adjudication, and Sentencing Outcomes
Editors-in-Chiefs Brian L. Cutler and Patricia A. Zapf

2015. 2,447 pages. Hardcover.

Series: APA Handbooks in Psychology®

List: \$395.00 | APA Member/Affiliate: \$195.00
ISBN 978-1-4338-1793-9 | Item # 4311515

APA Handbook of Career Intervention

Volume 1: Foundations
Volume 2: Applications
Editors-in-Chief Paul J. Hartung, Mark L. Savickas, and W. Bruce Walsh

2015. 1,008 pages. Hardcover.

Series: APA Handbooks in Psychology®

List: \$395.00 | APA Member/Affiliate: \$195.00
ISBN 978-1-4338-1753-3 | Item # 4311514

An APA LifeTools® Book

Write It Up Practical Strategies for Writing and Publishing Journal Articles

Paul J. Silvia, PhD
2015. 224 pages. Paperback.
List: \$19.95 | APA Member/Affiliate: \$19.95
ISBN 978-1-4338-1814-1 | Item # 4441024

How to Publish High-Quality Research Discovering, Building, and Sharing the Contribution

Jeff Joireman and Paul A. M. Van Lange
2015. 344 pages. Paperback.
List: \$29.95 | APA Member/Affiliate: \$24.95
ISBN 978-1-4338-1861-5 | Item # 4313037

Gestalt Therapy

Gordon Wheeler and Lena S. Axelsson
2015. 167 pages. Paperback.

Series: Theories of Psychotherapy Series®

List: \$24.95 | APA Member/Affiliate: \$24.95
ISBN 978-1-4338-1859-2 | Item # 4317359

Case Formulation in Emotion-Focused Therapy

Co-Creating Clinical Maps for Change
Rhonda N. Goldman and Leslie S. Greenberg
2015. 240 pages. Hardcover.
List: \$59.95 | APA Member/Affiliate: \$49.95
ISBN 978-1-4338-1820-2 | Item # 4317355

Spiritually Oriented Psychotherapy for Trauma

Edited by Donald F. Walker, Christine A. Courtois, and Jamie D. Aten
2015. 292 pages. Hardcover.
List: \$59.95 | APA Member/Affiliate: \$49.95
ISBN 978-1-4338-1816-5 | Item # 4317354

Treatment of Late-Life Depression, Anxiety, Trauma, and Substance Abuse

Edited by Patricia A. Areán
2015. 264 pages. Hardcover.
List: \$59.95 | APA Member/Affiliate: \$49.95
ISBN 978-1-4338-1839-4 | Item # 4317357

Men's Gender Role Conflict

Psychological Costs, Consequences, and an Agenda for Change
James M. O'Neill
2015. 400 pages. Hardcover.
List: \$79.95 | APA Member/Affiliate: \$49.95
ISBN 978-1-4338-1818-9 | Item # 4318128

The Lives of LGBT Older Adults

Understanding Challenges and Resilience
Edited by Nancy A. Orel and Christine A. Fruhauf
2015. 256 pages. Hardcover.
List: \$69.95 | APA Member/Affiliate: \$49.95
ISBN 978-1-4338-1763-2 | Item # 4318127

Interdisciplinary Frameworks for Schools

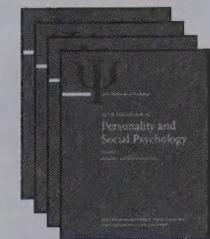
Best Professional Practices for Serving the Needs of All Students
Virginia Wise Berninger
2015. 432 pages. Hardcover.
List: \$79.95 | APA Member/Affiliate: \$49.95
ISBN 978-1-4338-1808-0 | Item # 4317352

Testing Accommodations for Students With Disabilities

Research-Based Practice
Benjamin J. Lovett and Lawrence J. Lewandowski
2015. 304 pages. Hardcover.
Series: Division 16: School Psychology Book Series
List: \$69.95 | APA Member/Affiliate: \$49.95
ISBN 978-1-4338-1797-7 | Item # 4317348

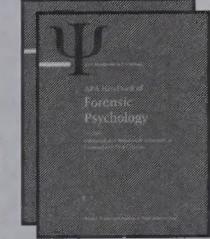
Forgiveness Therapy

An Empirical Guide for Resolving Anger and Restoring Hope
Second Edition
Robert D. Enright and Richard P. Fitzgibbons
2015. 352 pages. Hardcover.
List: \$69.95 | APA Member/Affiliate: \$49.95
ISBN 978-1-4338-1837-0 | Item # 4317358



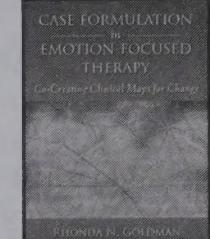
A Practical Guide to PTSD Treatment

Pharmacological and Psychotherapeutic Approaches
Edited by Nancy C. Bernady and Matthew J. Friedman
2015. 192 pages. Paperback.
List: \$29.95 | APA Member/Affiliate: \$29.95
ISBN 978-1-4338-1832-5 | Item # 4317356



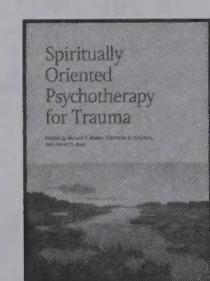
Psychological Practice With Women

Guidelines, Diversity, Empowerment
Edited by Carolyn Zerbe Enns, Joy K. Rice, and Roberta L. Nutt
2015. 304 pages. Hardcover.
Series: Division 35: Psychology of Women
List: \$69.95 | APA Member/Affiliate: \$49.95
ISBN 978-1-4338-1812-7 | Item # 4317353



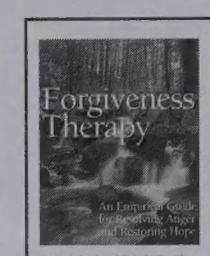
Prevention Psychology

Enhancing Personal and Social Well-Being
John L. Romano
2015. 216 pages. Hardcover.
List: \$69.95 | APA Member/Affiliate: \$49.95
ISBN 978-1-4338-1791-5 | Item # 4317347



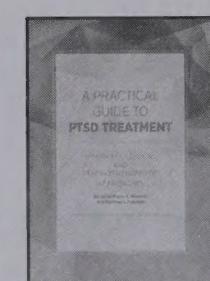
Biopsychosocial Practice

A Science-Based Framework for Behavioral Health Care
Timothy P. Melchert
2015. 352 pages. Hardcover.
List: \$69.95 | APA Member/Affiliate: \$49.95
ISBN 978-1-4338-1761-8 | Item # 4317346



Premature Termination in Psychotherapy

Strategies for Engaging Clients and Improving Outcomes
Joshua K. Swift and Roger P. Greenberg
2015. 216 pages. Hardcover.
List: \$69.95 | APA Member/Affiliate: \$49.95
ISBN 978-1-4338-1801-1 | Item # 4317349



Graduate Study in Psychology

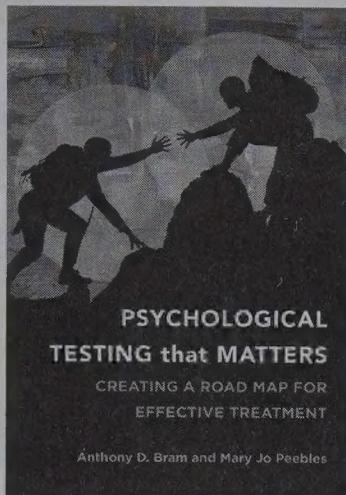
2015 Edition
2015. 976 pages. Paperback.
List: \$29.95 | APA Member/Affiliate: \$24.95
ISBN 978-1-4338-1780-9 | Item # 4270099

AVAILABLE ON AMAZON KINDLE®

PSYCHOLOGICAL TESTING THAT MATTERS

Creating a Road Map for Effective Treatment

Anthony D. Bram and Mary Jo Peebles



Psychological testing is widespread today. Test results are only valuable, though, when they contribute meaningful information that helps therapists better meet the treatment needs of their clients. *Psychological Testing that Matters* describes an approach to inference-making and synthesizing data that creates effective, individualized treatment plans. The book's treatment-centered approach describes how to reconcile the results of various tests, use test results to assess a patient's psychological capacities, reach a diagnosis, and write an informative test report.

2014. 464 pages. Hardcover.

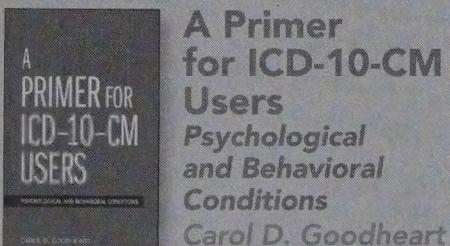
List: \$79.95 | APA Member/Affiliate: \$59.95 | ISBN 978-1-4338-1674-1 | Item # 4317333

CONTENTS:

Introduction | Section I: Basic Framework | Chapter 1. Treatment-Centered Diagnosis and the Role of Testing | Chapter 2. Inference Making | Chapter 3. Test Referral and Administration | **Section II: Key Psychological Capacities to Assess and Where to Look in the Data** | Chapter 4. Reasoning and Reality Testing | Chapter 5. Emotional Regulation: Balance and Effectiveness | Chapter 6. Experience of Self and Other Part A: Implications for Alliance | Chapter 7. Experience of Self and Other Part B: Narcissistic Vulnerability | **Section III: Diagnostic Considerations** | Chapter 8. Conceptualizing Underlying Developmental Disruption | Chapter 9. Assessing Underlying Developmental Disruption: Case Illustrations | **Section IV: Putting It All Together** | Chapter 10. Communicating our Findings: Test Report Writing and Feedback | Chapter 11. Detailed Case Example with Sample Report

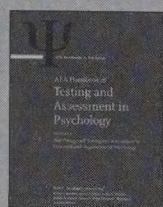
ALSO OF INTEREST

AVAILABLE ON AMAZON KINDLE®



2014. 200 pages. Spiral Binding.

List: \$19.95 | APA Member/Affiliate: \$14.95
ISBN 978-1-4338-1709-0 | Item # 4317336



2013. 2,010 pages. Hardcover.

• Series: APA Handbooks in Psychology®

List: \$695.00 | APA Member/Affiliate: \$395.00
ISBN 978-1-4338-1227-9 | Item # 4311510

APA Handbook of Testing and Assessment in Psychology
THREE VOLUME SET
Editor-in-Chief
Kurt F. Geisinger



Lorraine D. Eyde, Gary J. Robertson, and Samuel E. Krug
2010. 217 pages. Paperback.

List: \$19.95 | APA Member/Affiliate: \$19.95
ISBN 978-1-4338-0556-1 | Item # 4311013

Responsible Test Use
Case Studies for Assessing Human Behavior
SECOND EDITION

APA BOOKS ORDERING INFORMATION: 800-374-2721 • www.apa.org/pubs/books

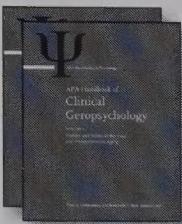
In Washington, DC, call: 202-336-5510 • TDD/TTY: 202-336-6123 • Fax: 202-336-5502
In Europe, Africa, or the Middle East, call: +44 (0) 1767 604972



AMERICAN PSYCHOLOGICAL ASSOCIATION

NEW RELEASES

from the American Psychological Association



APA Handbook of Clinical Geropsychology

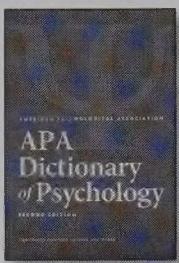
Volume 1: History and Status of the Field and Perspectives on Aging
Volume 2: Assessment, Treatment, and Issues of Later Life

Editors-in-Chief Peter A. Lichtenberg and Benjamin T. Mast

2015. 1,424 pages. Hardcover.

• Series: APA Handbooks in Psychology®

List: \$395.00 | APA Member/Affiliate: \$195.00
ISBN 978-1-4338-1804-2 | Item # 4311516



APA Dictionary of Psychology

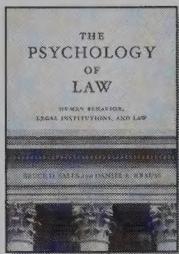
SECOND EDITION

Editor-in-Chief:

Gary R. VandenBos

2015. 1,2014 pages. Hardcover

List: \$49.95 | APA Member/Affiliate: \$39.95
ISBN 978-1-4338-1944-5 | Item # 4311022



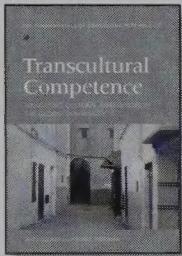
The Psychology of Law

Conceptualizing the Field for the 21st Century

Bruce D. Sales and Daniel A. Krauss

2015. 200 pages. Hardcover.

List: \$69.95 | APA Member/Affiliate: \$49.95
ISBN 978-1-4338-1936-0 | Item # 4316165

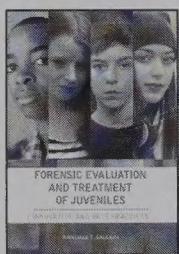


Transcultural Competence

Navigating Cultural Differences in the Global Community

Jerry Glover and Harris L. Friedman
2015. 176 pages. Paperback.

List: \$34.95 | APA Member/Affiliate: \$29.95
ISBN 978-1-4338-1945-2 | Item # 4317366



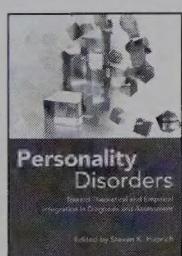
Forensic Evaluation and Treatment of Juveniles

Innovation and Best Practice

Randall T. Salekin

2015. 264 pages. Hardcover.

List: \$69.95 | APA Member/Affiliate: \$49.95
ISBN 978-1-4338-1934-6 | Item # 4317364

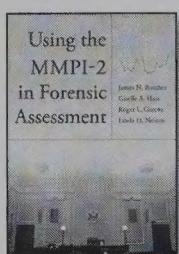


Personality Disorders

Toward Theoretical and Empirical Integration in Diagnosis and Assessment

Steven K. Huprich
2015. 440 pages. Hardcover.

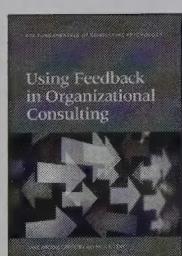
List: \$79.95 | APA Member/Affiliate: \$59.95
ISBN 978-1-4338-1845-5 | Item # 4316164



Using the MMPI-2 in Forensic Assessment

James N. Butcher, Giselle A. Hass, Roger L. Greene, and Linda D. Nelson
2015. 352 pages. Hardcover.

List: \$79.95 | APA Member/Affiliate: \$54.95
ISBN 978-1-4338-1868-4 | Item # 4317362



Using Feedback in Organizational Consulting

Jane Brodie Gregory and Paul E. Levy
2015. 166 pages. Paperback.

List: \$34.95 | APA Member/Affiliate: \$29.95
ISBN 978-1-4338-1951-3 | Item # 4317367



Violent Offenders

Appraising and Managing Risk

THIRD EDITION

Grant T. Harris, Marnie E. Rice, Vernon L. Quinsey, and Catherine A. Cormier
2015. 480 pages. Hardcover.

List: \$79.95 | APA Member/Affiliate: \$49.95
ISBN 978-1-4338-1901-8 | Item # 4317363



AMERICAN PSYCHOLOGICAL ASSOCIATION

TO ORDER: 800-374-2721 • www.apa.org/pubs/books

AD2812

STATA does more.

One unified statistical software program for all of your analytical needs.

The ease of a user-friendly interface.

The flexibility of a matrix programming language.

Stata's clean interface is arranged to simplify your workflow. The Data Editor, Graph Editor, and dialog boxes ease all types of analyses. But there are no restrictions. With Stata's intuitive command syntax and matrix programming language, you have the freedom to customize Stata to perfectly suit your needs.

ANOVA, CFA, hierarchical models, growth curves, interaction plots, SEM ... Stata does all this, and more.

STATA®



stata.com/edu15

Stata is a registered trademark of StataCorp LP, 4905 Lakeway Drive, College Station, TX 77845, USA.